EDITED BY

HERMAN CAPPELEN

TAMAR SZABÓ GENDLER

JOHN HAWTHORNE

The Oxford Handbook *of* PHILOSOPHICAL METHODOLOGY

THE OXFORD HANDBOOK OF

PHILOSOPHICAL METHODOLOGY

THE OXFORD HANDBOOK OF

PHILOSOPHICAL METHODOLOGY

Edited by

HERMAN CAPPELEN

TAMAR SZABÓ GENDLER

and

JOHN HAWTHORNE

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data

Data available

Library of Congress Control Number: 2016933497

ISBN 978–0–19–966877–9

Printed in Great Britain by
Clays Ltd, St Ives plc

Preface

Philosophy is a field where methodology plays two roles: it is both a framework for, and a subject of, philosophical investigation. These two roles are deeply intertwined. Most significant periods and movements in philosophy have been—at least in part—instigated by a renewed methodology. And this renewal is typically accompanied by a renewed self-consciousness—an explicit conception of philosophy and philosophizing as themselves subjects of philosophical reflection. A surprisingly large portion of good philosophy is not only methodologically self-aware, but also methodologically innovative: think not only of Plato, Descartes, and Wittgenstein, but also of the ways in which contemporary strands in philosophy—from analytic metaphysics to feminist philosophy to empirically informed work in moral psychology—are simultaneously interventions into first-order and second-order philosophical questions. A volume on philosophical methodology is thus not merely a work about how philosophy is done, but also a work of philosophy.

Philosophical methodology is not a field that comes pre-structured. What you think counts as (a significant part of) methodology (or how you draw the distinction between first-and-second order philosophy) and will depend in large part on your philosophical views—in much the same way as what you count as philosophy does. That rendered the task of putting together a volume like this one both challenging and exciting: part of our task was simply figuring out the nature of the task. Many of our contributors asked us what *exactly* we meant by "methodology" and we were deliberately non-committal in our responses: We didn't want to impose a conception of methodology on the volume and our contributors. As a result, the contributions in this volume illustrate the wide range of conceptions of what counts as philosophical methodology. That is a central part of the volume's intention: a carefully curated tour of a range of meta-philosophical perspectives.

That said, the selection of topics does of course reveal something about our perspective. One feature of the volume is particularly important to highlight: we have put a great deal of focus on the connections between philosophy and other disciplines. About a third of the volume is devoted to explorations of these connections. This decision reflects what we think is an important and characteristic feature of philosophy, both in its contemporary incarnation and in its history. At its core, philosophy draws on, contributes to, and connects fields outside philosophy. Maybe more so than any other discipline, philosophy looks outside of itself. There is no limit on the kind of data and theorizing that a philosopher can draw on. There are, of course, philosophers who argue against this conception of philosophy and who think of philosophy as autonomous and fundamentally different from, for example, empirical disciplines. But, as we see it, for those arguments to be remotely plausible they have to be understood as normative claims about what philosophy ought to be. If we just look at the facts about what philosophers have done and are doing, then this openness to other disciplines is striking.

We should admit to a limitation of this volume. The entries all concern what we can coarsely describe as the "Western tradition" of philosophy, and are largely Anglo-American in approach. The volume is already massive (pushing the limits of what the publisher would accept), and adding a correspondingly full range of entries on other approaches was out of the question. The alternative—a single 20-page entry exploring themes from different traditions in rapid-fire fashion—would fail to do justice to their significance and complexity.

Even within our hemisphere, there is a rich world to explore. The volume begins with a meditation on its central question: What is philosophical methodology? This is followed by a cluster of essays that we call "Traditions and Approaches" which examine the topic of philosophical methodology in the Western philosophical tradition from Ancient Greece to the modern anglophone world. The next section explores a range of central "topics" in philosophical methodology—from reflective equilibrium to conceptual analysis to philosophical heuristics. The final section—"Philosophy and its Neighbors"—explores the boundaries (or non-boundaries) between philosophical investigation and investigation in fields such as linguistics, psychology, logic, and political theory.

Though we encouraged authors to read and respond to other essays in the volume, each of the essays has retained the "voice" of its author. This, too, was a deliberate decision. The range of methodological perspectives that this volume represents deserves this range of minds that have been brought together to introduce them. We hope that you will enjoy reading these marvelous entries as much as we have enjoyed preparing them for you.

Herman Cappelen
Tamar Szabó Gendler
John Hawthorne

Acknowledgements

We are grateful to our three research assistants, Joshua Habgood-Coote, Aaron Norby, and Daniele Sgaravatti, for invaluable support throughout the project.

Contents

PART III TOPICS

PART IV PHILOSOPHY AND ITS NEIGHBOURS

Figures

About the Contributors

Louise Antony, University of Massachusetts

Avner Baz, Tufts University

Frederick C. Beiser, Syracuse University

John P. Burgess, Princeton University

Taylor Carman, Barnard College

Yuri Cath, La Trobe University

Gregory Currie, University of York

Josh Dever, University of Texas at Austin

Paul Franks, Yale University

Gary Gutting, University of Notre Dame

Alan Hájek, Australian National University

Christopher Hill, Brown University

Paul Horwich, New York University

Henry Jackman, York University

Jeffrey C. King, Rutgers

Thomas Kelly, Princeton University

Hilary Kornblith, University of Massachusetts

Alex Langlinais, Yale University

Brian Leiter, University of Chicago

Christian List, London School of Economics

Dominic McIver Lopes, University of British Columbia

Peter Ludlow, Northwestern University

Ishani Maitra, University of Michigan

Ron Mallon, Washington University in St. Louis

Charles Mills, CUNY Graduate Center

Daniel Nolan, Australian National University

Calvin G. Normore, McGill University

Derk Pereboom, Cornell University

Christopher Pincock, Ohio State University

Georges Rey, University of Maryland

Adina L. Roskies, Dartmouth College

Laura Ruetsche, University of Michigan

Stewart Shapiro, Ohio State University

Scott Soames, University of Southern California

Laura Valentini, London School of Economics

Brian Weatherson, University of Michigan

Jonathan M. Weinberg, University of Arizona

Michael Weisberg, University of Pennsylvania

Linda Zagzebski, University of Oklahoma

PART I

BACKGROUND

CHAPTER 1

WHAT IS PHILOSOPHICAL METHODOLOGY?

JOSH DEVER

1. Introduction

Philosophical methodology has been a hot topic recently. (I hear they're devoting entire *handbooks* to it!) But what is the topic? What do we talk about when we talk about philosophical methodology? The goal of this chapter is to address that *meta*methodological question.[1] I won't attempt to say what good philosophical methodology is, or to answer any first-order methodological questions, but only to figure out what kind of thing a philosophical methodology (good or not) would be, and what kind of questions would count as methodological questions.

A persistent ambiguity dogs any discussion of the metamethodological question of what kind of thing philosophical methodology is, rather than (e.g.) what instances of that type are typically or ideally realized. The question "What is philosophical methodology?" admits of a "lower-order" reading, on which admissible answers are *the use of thought experiments to test conceptual analyses*, or *understanding us and our environments in a way conducive to human flourishing*. But the same question admits of a "higher-order" reading, on which admissible answers are *the epistemological methods that distinguish philosophy from the natural sciences (on one side) and the humanities (on the other side)*, or *the pursuit of a description of reality at the most fundamental level*. It is the "higher-order" reading that I will be primarily interested in in this chapter, but to avoid ambiguity or the need for lengthy disambiguations, I will use "Philosophical Methodology" to pick out questions

[1] I use "meta" here in a loosely type-theoretic sense, so that metamethodological questions are questions that refer to something referring to methodology, rather than merely questions that refer to methodology. The use of the "meta"-label is also a way of characterizing the subfield of philosophical methodology—it is presumably not an orthographic accident that it is often called "metaphilosophy." The current topic, in that terminology, is metametaphilosophy. The question of the relation between the two occurrences of "meta" in this label is taken up in section 5.

of the higher order, and "philosophical methodology" to pick out questions of the lower order.[2]

It is perhaps appropriate to begin this investigation into the nature of Philosophical Methodology with a brief description of the methodology that the investigation itself will use.[3] We'll start by treating this as a problem in (not so) ordinary language philosophy: there is a term "philosophical methodology" at use in philosophical talk and writing, and we want to know what that term picks out. To provide a robust data pool, I'll take all occurrences of the word "methodological" in entries of the *Stanford Encyclopedia of Philosophy (SEP)*, and see what plausible theories of Philosophical Methodology can be fitted onto the bed of that range of usage.[4]

We will turn shortly to the consideration of some theories of Philosophical Methodology on the basis of this data pool, but let's begin by looking at a few representative usages to get a sense for the shape and difficulty of the task. Here are five samples from the data pool, selected (from among many other suitable candidates) to display some of the paradigm usage types:

1. [Aquinas] employs, through all his works, a **methodological** axiom: *X*'s nature is understood by understanding *X*'s capacities, which are understood by understanding their act[uation]s, which are understood by understanding their objects. (John Finnis, *SEP*, "Natural Law Theories")[5]
2. From this vantage point, general covariance is but the most recent refinement of the **methodological** principle of "unity of determination" governing the constitution of objects of physical knowledge, completing the transposition in physics from concepts of substance into functional and relational concepts. (Thomas Ryckman, *SEP*, "Early Philosophical Interpretations of General Relativity")
3. Is there any a priori support for "There is nothing"? One might respond with a **methodological** principle that propels the empty world to the top of the agenda. For instance, many feel that whoever asserts the existence of something has the burden of proof. (Roy Sorensen, *SEP*, "Nothingness")

[2] So "What philosophical methodology should we use to determine what Philosophical Methodology is?" is the question of what member of some type T we should use in coming to understand the nature of type T itself. The lower-order questions will typically be more amenable to normative instances, because the metaphysical/conceptual questions of the nature of the category are less normative (which is not to say wholly non-normative) than questions about the use of members of the category.

[3] Is it also perhaps incoherent, on the grounds that we are in no position to consider a particular methodology until we've determined what Methodology is? No, for familiar *Meno*-type reasons. It suffices that we have a recognitional capacity with respect to methods. So long as you recognize what I give as a method, that will suffice to begin the process of determining what kind of thing you are thereby recognizing it as being.

[4] A brief summary of the data: there are (as of 2010, when the original data was collected) 285 entries in the *Stanford Encyclopedia of Philosophy* that use the word "methodological." The full list of those entries (in the order produced by the search engine) is shown in Appendix 1 of this chapter.

[5] Bolding of "methodological" and cognates is added to quotations throughout. All other forms of emphasis are in the original.

4. Aristotle approaches his account of the nutritive soul by relying on a **methodological** precept which informs much of his psychological theorizing, namely that a capacity is individuated by its objects, so that, e.g. perception is distinguished from mind by being arrayed toward sensible qualities rather than intelligible forms. (Christopher Shields, *SEP*, "Aristotle's Psychology")
5. Cognitive science raises many interesting **methodological** questions that are worthy of investigation by philosophers of science. What is the nature of representation? What role do computational models play in the development of cognitive theories? What is the relation among apparently competing accounts of mind involving symbolic processing, neural networks, and dynamical systems? What is the relation among the various fields of cognitive science such as psychology, linguistics, and neuroscience? Are psychological phenomena subject to reductionist explanations via neuroscience? (Paul Thagard, *SEP*, "Cognitive Science")

Whatever Philosophical Methodology is, it is apparently the source of axioms, principles, precepts, and questions. Making the bold assumption that axioms, principles, and precepts are all the same thing, we have a short list of propositions wearing the "methodological" honorific:

- (A) X's nature is understood by X's capacities, (B) "For the physical description of the processes of nature no particular reference body should be distinguished above all others",[6] (C) Whoever asserts the existence of something has the burden of proof, (D) A capacity is individuated by its objects

We also have a list of questions distinguished as "methodological", including "What role do computational models play in the development of cognitive theories?" and "Are psychological phenomena subject to reductionist explanations via neuroscience?".

Even the small sample of propositions (A) through (D) shows a considerable diversity. (D) is a straightforward proposition of metaphysics and (C) a straightforward proposition of epistemology. (A) lands either in metaphysics or in epistemology, depending on whether "understood" is given a metaphysical (e.g. grounding) or an epistemological (e.g. explanatory) reading. (B) is a normative claim, in which the relevant norm is something like a (non-alethic?) norm of theory selection. The additional methodological questions rapidly add to the diversity. It is tempting, in order to constrain the class of the Methodological, to hypothesize that a methodological question is not a methodological question simply in virtue of being a question whose answer is a methodological proposition (=axiom, principle, etc.), but rather that it is somehow distinctively the *question* that is methodological. At this early point in the investigation, however, it is not transparent how to make this distinction.

What, then, *do* we talk about when we talk about philosophical methodology? I think even this small sample of the usage suffices to show that there's no obvious and straightforward fitting of an account of Philosophical Methodology onto the way in which

[6] The "principle of the unity of determination", from Cassirer (1921), as discussed in Ryckman (2005).

philosophers use "methodological" talk.[7] In the remainder of this investigation I'll set out a sequence of attempted fittings, which will be tested against a wider range of the data as we proceed. More or less each attempt will try to respond to the shortcomings of the previous attempt, so if all goes well there will be some accumulation of virtues along the way.

It's a long journey through the sequence of attempts, though, so I'll start with a math problem to mull over if you weary of Methodological hypotheses:

The Puzzle: Springfield is a town of 1,000 people, whose residents have a mania for forming clubs. The residents of Springfield place two requirements on their clubs:

1. Each club must have an odd number of members.
2. Any two distinct clubs must have an even number (possibly 0) of members in common.

Determine (with proof) the largest number of clubs that the residents of Springfield can form.

2. Hypothesis #1: Eliminativism

Let's begin with the simplest view: there's nothing at all meant by "methodological" in philosophical contexts. It can simply be globally eliminated without any content being lost.[8] To see why one might be tempted by Eliminativism, consider claims such as:

- A simple scientific example can be found in the rationale behind the sensible **methodological** adage that "correlation does not imply causation." (Kyle Stanford, *SEP*, "Underdetermination of Scientific Theory")
- There is, then, at the very least, a **methodological** issue separating DRT from non-representational frameworks: a non-representational framework is defined with a natural criterion for identity of meaning of two sentences in mind, and this criterion is related to the information conveyed by those sentences. But in a representational framework like DRT the representations themselves provide the only criterion for judging identity of meaning. So is representationalism only **methodological**, a mere convenience? (Bart Geurts and David Beaver, *SEP*, "Discourse Representation Theory")

[7] One hypothesis is that the subfield of "Philosophical Methodology" is not, in fact, beholden to how philosophers at large use the "methodological" vocabulary, but rather only to how philosophers *engaged in the 'Philosophical Methodology' subfield* use the "methodological" vocabulary. (There is a worry here about how to identify the relevant subfield, but we might get started by ostension, or by characterizing a distinctive "meta" syntactic profile that distinguishes talk about Philosophical Methodology in the relevant sense.) I won't pursue this hypothesis here, but plausibly it would lead (for reasons that are alluded to in section 4) to something similar to what I call "Epistemologism" about Philosophical Methodology.

[8] Plausibly Eliminativism should be the null hypothesis for all philosophical vocabulary.

Now re-read those claims deleting the occurrences of "methodological."[9] How much is lost by taking "causation does not imply correlation" to be merely an adage, without specifying that it is a *methodological* adage, or by taking the difference in criterion for the identity of meaning to be an issue separating representational from non-representational semantic frameworks, without specifying that it is a *methodological* issue?

Or, to venture outside the bounds of the *Stanford Encyclopedia of Philosophy*, consider Bonevac, Dever, and Sosa (2011)'s "First Methodological Observation":

- Counterexamples must be deployed as counterexamples to *specific proposals*. The example of a glass packed in Styrofoam can perhaps show that *fragile* cannot be analyzed as *would break if struck*, but it shows nothing about a proposed analysis of *fragile* as *would break if struck when unwrapped*, and *certainly* shows nothing about any proposed analysis of a different dispositional term, such as *irascible*.

Again, it's not clear that anything is gained by calling this a *methodological* observation, rather than just an observation, beyond a certain weightiness of style.[10]

Eliminativism about Philosophical Methodology can't be taken too seriously as a proposal. Methodological talk is widespread throughout philosophy, and Eliminativism would require a rather stark error theory about our stance toward our own philosophical practice. What *does* emerge from the consideration of Eliminativism is that, while there is surely a robust and substantive notion of Philosophical Methodology embedded in our practice, it also looks undeniable that philosophical rhetoric has an inclination toward "methodological creep"—there is a temptation to achieve *gravitas* by calling points methodological.[11] It's thus at least a minor desideratum for a theory of Philosophical Methodology that it explain why "methodological" is, as it were, a high status word (at least among philosophers).

3. Hypothesis #2: Working-Hypothesism

Consider the use of "methodological" language in the discussion of methodological reductionism in Ingo Brigandt and Alan Love's *SEP* entry on reductionism in biology:

- ***Methodological** reduction* is the idea that biological systems are most fruitfully investigated at the lowest possible level, and that experimental studies should be aimed at uncovering molecular and biochemical causes. A common characterization of this

[9] And making appropriate syntactic adjustments.

[10] A certain amount of uncomfortable autobiographical reflection was involved in the writing of this piece, as will manifest again in this chapter.

[11] Of course, the apparent eliminability of "methodological" from any particular bit of philosophical text is inadequate to determine that that usage is rhetorical, rather than substantive. Even if nothing in the content of a given text *requires* or *determines* a non-vacuous meaning of "methodological", the text can still be exploiting a substantive conception of Philosophical Methodology determined by other parts of the practice.

> type of strategy is what has been termed "decomposition and localization" (Bechtel and Richardson 1993). While **methodological** reductionism is often motivated by the presumption of ontological reduction, this procedural recommendation does not follow directly from it. (Brigandt and Love, *SEP*, "Reductionism in Biology")

Eliminitivism is not a plausible analysis of this use of "methodological"—because methodological reductionism is being contrasted with other forms of reductionism, some substantive theory of what makes a form of reductionism peculiarly *methodological* is needed.

Here's a crude diagnosis: we focus on Brigandt and Love's claim that methodological reductionism is a technique for securing *fruitful* biological theorizing by, in effect, *acting as if* (biological) reductionism is true, whether or not it is in fact true. We see a similar idea in Zoltán Szabó's *SEP* article on compositionality. Here, under the subsection "Methodology", Szabó says:

- By far the most popular reason for believing in compositionality is that it works. Linguists have adopted various versions of the principle as a working hypothesis and developed semantic theories on their basis. These theories have provided intuitively satisfactory explanations for certain data, such as the validity or invalidity of certain inferences or for various sorts of contrasts between certain minimal pairs. (Szabó, *SEP*, "Compositionality")

Like Brigandt and Love, Szabó characterizes a principle as methodological on the grounds that guiding theorizing by it produces better final theories. (Szabó differs from Brigandt and Love by characterizing the theoretical fruitfulness of the principle as something people take as a reason for belief in the principle, although he proceeds to say that "despite its popularity, this is not a very good reason to believe in compositionality", and most of his methodological discussion is independent of the question of outright belief in the methodological principle.)

Philosophical Methodology is, on this conception, the task of identifying "working hypotheses", tentative assumptions, and norms of investigation that lead to good theorizing. Working hypotheses in the relevant sense can be claims that we suspect, but don't yet know, to be true, such as the principle of compositionality in semantic theorizing. Here the pursuit of compositional meaning theories is perhaps fruitful in part because (we suspect) the one true semantic theory is in fact compositional, so our working hypothesis directs us to the proper part of logical space. But the compositional pursuit can also be fruitful because it helps us avoid certain kinds of mistakes by lowering the complexity of the theory, or because it makes it easier for us to integrate modular theories for different semantic phenomena. Working hypotheses can also be claims that are (we suspect) false, such as a claim of ontological reductionism in biology. In adopting reductionism as a working hypothesis, we perhaps obtain better theories, not because the biological systems are genuinely reductive, but rather because there are de facto widespread explanatory relations, not sufficient for reductionism, from the microphenomena to the macrophenomena.

The relevant tentative assumptions can, it would appear, be arbitrarily diverse in their content—they need not be, for example, exclusively "deep" metaphysical or epistemological principles. "Working hypotheses" in this sense can also cover simple advice on

how to do good philosophical theorizing, as in Bonevac, Dever, and Sosa (2011)'s "Third Methodological Observation":

- Philosophical discussions of conditionals should not assume that the logical options are limited to the material condition, the C. I. Lewis strict conditional, and the Lewis/Stalnaker variably strict counterfactual conditional.[12]

But if Philosophical Methodology amounts simply to nuggets of advice on how to get good philosophy done, there is a worry about how to carve out a philosophically interesting enterprise under the "Philosophical Methodology" banner, rather than a glorified anthology of avuncular wisdom. There is no clear stopping point on the spectrum from "assume until forced to do otherwise that biological phenomena have reductive biochemical explanations" to "try to make your semantic theories compositional, unless that gets hopelessly messy", to "consider a wide range of logical options", to "don't do philosophy while drunk".[13]

4. Hypothesis #3: Epistemologism

Many instances of "methodological" talk focus on epistemological questions of how we achieve *knowledge* through philosophical activity. Consider some representative samples:

- However it is often thought that knowledge of animal minds—what Allen & Bekoff (1997) refer to as "the other species of mind problem" and Prinz (2005) calls "The Who Problem"—presents special **methodological** difficulties because we cannot interrogate animals directly about their experiences (but see Sober 2000 for discussion of tractability within an evolutionary framework, and Farah 2008 for a neuroscientist's perspective). Although there have been attempts to teach human-like languages to members of other species, none has reached a level of conversational ability that would solve this problem directly (see Anderson 2004 for a review). (Colin Allen, *SEP*, "Animal Consciousness")
- Ironically, the identification of empathy and understanding and the associated claim that empathy is the sole and unique **method** of the human sciences also facilitated the decline of the empathy concept and its almost utter disregard by philosophers of the human and social sciences later on, in both the analytic and continental/hermeneutic traditions of philosophy. Within both traditions, proponents of empathy were—for very different reasons—generally seen as advocating an epistemically naïve and insufficiently broad conception of the **methodological** proceedings in the human sciences. As a result, most philosophers of the human and social sciences maintained their distance from the idea that empathy is central for our understanding of other minds and mental phenomena. (Karsten Stueber, *SEP*, "Empathy")

[12] A certain amount of uncomfortable autobiographical reflection is reflected here also.
[13] Or "don't do philosophy while sober", if that's your preference. I'm not here to judge.

- Informed by his reading of Schleiermacher, Droysen, and Dilthey, Martin Heidegger's *Sein und Zeit* (1927) completely transformed the discipline of hermeneutics. In Heidegger's view, hermeneutics is not a matter of understanding linguistic communication. Nor is it about providing a **methodological** basis for the human sciences. (Bjørn Ramberg and Kristin Gjesdal, *SEP*, "Hermeneutics")

The difficulties created by animal minds are specifically *methodological* difficulties because they are epistemological difficulties—the problems here are not, for example, problems in accounting for the metaphysics of animal minds, but for our knowledge of them. Similarly, empathy is a *methodological* issue because it is (in context) an epistemic tool for reaching conclusions in the human sciences. The general hypothesis, then, is that Philosophical Methodology is the study of the means by which we come to achieve knowledge[14] in philosophy.

On the assumption that what we are trying to do when we do philosophy is to come to know various (philosophical) truths about the world, Epistemologism about Philosophical Methodology sweeps in the "working hypothesis" view as a special case. Methods as working hypotheses are tools to help us do philosophy better, which then translates into tools for achieving knowledge. But the "methodological tidbits" considered in section 3 will represent ad hoc tinkering about the edges of the problem, while the central topic will be a thorough characterization of our epistemic relation to philosophical claims.

Epistemologism about Philosophical Methodology encompasses much of the recent work that goes under the specific banner of "philosophical methodology." We can distinguish two classes of issues that emerge in Epistemologist Philosophical Methodology. There are *descriptive* issues, centering on a specification of the epistemic methods that are conducive to philosophical knowledge. And there are *normative* issues, centering on the question of why the methods specified when addressing the descriptive question are the right methods to use.[15]

Recent discussion of the descriptive issues has revolved around renewed interest in characterizing philosophical investigation. One persistent topic has been the role of thought experiments as a tool for testing philosophical theses. The descriptive questions here are numerous. We need to know what kind of philosophical thesis can be tested by thought experiments (only necessary theses?), what kinds of things thought experiments are (for example, what the *content* of a thought experiment is), whether thought experiments are a global epistemic device across philosophy or whether their application is limited to specific subdomains of philosophy, whether the reactions of all subjects to thought experiments are

[14] Or justification, or etc. We needn't fuss about the precise epistemological category here.

[15] The labeling here is tricky, because the epistemic methods being enumerated and described in addressing the descriptive question are themselves normative methods, or at least intimately related to normative claims. Answering the descriptive question will lead us to say, for example, that intuitions **justify** philosophical claims, or that experimental methods **ought** to be used to check judgments on thought experiments, or that **good** theorizing is done through conceptual analysis. Answering the normative question leads us to say that intuitions justify because intellectual seemings, like perceptual seemings, grant default entitlement, or that conceptual analysis grounds apparent synthetic a priori knowledge via our ability to create truth through stipulation. We might equally well call the two categories the *normative* and the *metanormative*.

equally valid or whether successful/epistemically weighty evaluation of a thought experiment requires specific skills or expertise, whether evaluation of a thought experiment is underwritten by specific epistemic faculties such as intuition or a generic capacity for the evaluation of counterfactuals, and so on.

More generally, the descriptive issues in Epistemologist Philosophical Methodology tend to blend seamlessly into general epistemological questions, especially in the epistemology of a priori knowledge. Since many of the propositions we want to assess in philosophy appear to be either a priori true or a priori false (although the extent of the involvement of *a prioricity* in philosophical practice is another topic in Epistemologist Philosophical Methodology), describing the epistemic methods of philosophy will largely involve describing the methods of achieving a priori knowledge. Questions about the role of intuitions, deductive reasoning, and conceptual analysis all reside in this aspect of the topic.[16]

The normative issues in Epistemologist Philosophical Methodology then flow from the descriptive issues. If intuitions play a crucial role in our philosophical method, what could the epistemology of intuitions be, such that they could be a source of knowledge? If conceptual analysis underwrites our philosophical knowledge, how does conceptual mastery translate into justification for analytic truths? And so on.

There is undeniably a rich and robust philosophical enterprise being described by Epistemologist Philosophical Methodology. It is, however, unclear whether that enterprise is in any way distinct from epistemology *tout court*. The worry here is twofold. First, it is an open question whether there is anything distinctively *philosophical* in philosophical methodology, taken in this epistemological sense. *Maybe* the ways of coming to know philosophical truths are different from the ways of coming to know non-philosophical truths (either because some general ways of knowing are not applicable in the philosophical domain, or because some distinctively philosophical ways of knowing are not applicable outside philosophy). But certainly an examination of the actual practice of philosophers does not give grounds for optimism about this kind of philosophical exceptionalism. Actual philosophical practice involves pretty much every imaginable epistemic method: arguments deductive and inductive, experimentation, opinion survey, reflective equilibrium, critical self-examination, phenomenological introspection, textual exigesis and eisigesis, and so on. If there is nothing distinctively philosophical in philosophical methodology, and philosophical methodology is just the epistemology of philosophical practice, then philosophical methodology collapses into epistemology.

Second, even if there are distinctively and uniquely philosophical epistemic tools, the normative characterization about those tools may be continuous with the characterization of general epistemic tools. *Perhaps* evaluation of thought experiments is a distinctive philosophical epistemic method, but perhaps also the underlying account of our ability to

[16] Interest in these topics is not, of course, a novelty in philosophy. While there has certainly been a recent resurgence of writing on these sorts of methodological questions, attempts to characterize the epistemic methodology of philosophers appear recurrently throughout the historical record. To pick one of many possible examples, Russell extensively characterizes his method of analysis in *The Philosophy of Logical Atomism* and elsewhere. Stebbing's classic "The Method of Analysis in Metaphysics" makes important steps toward addressing the *normative* questions by asking, not just what the method of analysis is, but what the relation between that method and metaphysics could be, such that analysis would be a way of coming to know metaphysical truths.

achieve justified beliefs through consideration of thought experiments runs, Williamson-style, through "off-line" deployment of very general faculties for the representation and modeling of the world around us. If this is the case, then again philosophical methodology becomes indistinguishable from epistemology in general.

Not that there's anything wrong with epistemology. But there are aspects of the use of "methodology" language that are hard to square with the thought that Philosophical Methodology is just another name for epistemology. We'll turn to some specific recalcitrant data in section 5, but I'll close this section with a different sort of concern. One occasionally encounters criticisms of mainstream "analytic philosophy" for "methodological narrowness." Both continental and feminist critiques, for example, will at times raise such methodological concerns. Thus consider Nancy Tuana's "Approaches to Feminism" entry in the *SEP*, which says:

- There has been significant debate within feminist philosophical circles concerning the effectiveness of particular methods within philosophy for feminist goals. Some, for example, have found the **methods** of analytic philosophy to provide clarity of both form and argumentation not found in some schools of Continental philosophy, while others have argued that such alleged clarity comes at the expense of rhetorical styles and **methodological** approaches that provide insights into affective, psychic, or embodied components of human experience. Other feminists find approaches within American pragmatism to provide the clarity of form and argumentation sometimes missing in Continental approaches and the connection to real world concerns sometimes missing in analytic approaches. (Tuana, *SEP*, "Approaches to Feminism")

The concerns raised here, despite their "methodological" labeling, are not all manifestly epistemological concerns (consider, for example, the concern about "connection to real world concerns"). And if methodology just is epistemology, then the narrowness concerns (correct or incorrect) become thoroughly misguided—as noted above, the epistemic methods used by philosophers seem to be as broad as they could possibly be.[17] Charitable reconstruction of the narrowness concern, then, seems to require some non-epistemic dimension to Philosophical Methodology.

5. Hypothesis #4: Theory Selectionism

Sometimes, contra Epistemologism, "methodological" talk is explicitly *contrasted* with epistemological talk. Thus:

- Occam's Razor may be formulated as an *epistemic* principle: if theory *T* is simpler than theory *T**, then it is rational (other things being equal) to believe *T* rather than *T**. Or

[17] Some narrowness concerns can be construed as worries not about what the full range of deployed epistemic methods is, but about the relative frequencies with which various methods are used. But not all.

> it may be formulated as a ***methodological*** principle: if *T* is simpler than *T** then it is rational to adopt *T* as one's working theory for scientific purposes. (Alan Barker, *SEP*, "Simplicity")

Here to take Occam's Razor as a methodological principle is precisely *not* to take it as saying something about the role of simplicity considerations in our *epistemology*. As a rough first draft, we can distinguish between taking a realist stance toward our theories, which involves *believing* them, and taking an instrumentalist/anti-realist stance toward our theories, which involves merely *accepting* them.[18]

A view that takes Philosophical Methodology to be the principles that guide mere theory acceptance obviously contrasts sharply with Epistemologism, and will not account for the usages cited in section 4. Moreover, it will reduce Philosophical Methodology to a topic of interest only given a prior commitment to a particular philosophical position that asks for a belief/acceptance distinction in the first place. But we can find a characterization of Philosophical Methodology that allows unification of these two views. Consider another discussion of the role of simplicity in theorizing:

- Unity can be understood as a **methodological** principle (Wimsatt 1976 and Wimsatt 2006 for the case of biology and Cat 1998 for physics). One way of doing so is as a simplicity or parsimony condition. But this kind of condition can receive two different interpretations: epistemological and ontological. Sober has drawn this distinction to shed light on the methodological role of unity hypotheses. He formulates it as a distinction between Peirce's problem and Hempel's problem of deciding between competing hypotheses, one unified and the other disunified: Hempel's problem is how to choose between two descriptively true hypotheses in terms of explanatory value; Peirce's problem is how to choose between two explanations in the light of the data they explain in terms of their predictive accuracy or likelihood to be true. (Jodi Cat, *SEP*, "The Unity of Science")

Here again an epistemological and a non-epistemological use of simplicity is distinguished, but both are bundled under a common "methodological" heading. What makes them both "methodological", it would seem, is that, whether epistemological or not, both are being used as tools for theory selection.

Philosophical Methodology, on this view, is the general study of criteria for theory selection. X is a methodological axiom or principle if X is a consideration which favors selecting one theory over another; X is a methodological question if X is a question such that a

[18] This distinction, of course, has a long and, well-distinguished history in philosophy of science, running through Mach, Poincaré, Carnap and the Vienna Circle, and van Fraassen. I won't attempt here any non-caricatured presentation of the nuances of scientific anti-realism. What matters for our purposes is that the entanglement of "methodological" talk with the scientific realism/anti-realism dispute, as manifested in this strand of the usage of "methodological", seems to be responsible for the disproportionate representation of philosophy of science in the distribution of occurrences of "methodological" in the *SEP*. Although the historical data is sketchy, I think it's plausible that the specifically philosophy-of-science usage is the primary point of entry for "methodological" talk into the rest of philosophy.

theory's answer to that question plays a role in determining the acceptability of the that theory. Given that one desideratum for theory selection is that our selected theories be theories we are *justified* in endorsing, Epistemologism about Philosophical Methodology gets incorporated as a special case—facts about how we can achieve knowledge of philosophical claims turn into facts about how we can select one (philosophical) theory over another.

Criteria for theory selection can be global considerations (simplicity and unity, support by intuition or experiment, and so on). But they can also be local and specific to a particular theoretical enterprise. When we are told:

- In the search for a theory of quantum mechanics it became a **methodological** requirement to Bohr that any further theory of the atom should predict values in domains of high quantum numbers that should be a close approximation to the values of classical physics. (Jan Faye, *SEP*, "Copenhagen Interpretation of Quantum Mechanics")

we are thereby given a particular desideratum for theoretical adequacy (close approximation to classical physics in domains of high quantum numbers) that is *methodological* because it is part of the method for selecting a theory of quantum mechanics, but which is not part of a *general* account of theory selection.[19] Or again:

- Psychologist Zenon Pylyshyn (1984) appeals to multiple realizability to ground a **methodological** criticism of reductionism. He described a pedestrian, having just witnessed an automobile accident, rushing into a nearby phone booth and dialing a 9 and a 1. What will this person do next? Dial another 1, with overwhelming likelihood. Why? Because of a systematic generalization holding between what he recognized, his background knowledge, his resulting intentions, and that action (intentionally described). (John Bickle, *SEP*, "Multiple Realizability")

The methodological criticism of reductionism is that reductionism will make unavailable generalizations crucial to the prediction and explanation of human behavior. The criticism is thus that a reductionist theory of mind will inevitably fail to meet certain criteria for theoretical adequacy.

Philosophical Methodology is on this picture *metaphilosophy* in the sense that it is an attempt to characterize what we are trying to do when we do philosophy, rather that a first-order attempt to do philosophy. Again we can separate a descriptive and a normative aspect of the project. Normatively, the aim is to say what the goals of philosophy are (to give a complete catalog of truths, to give a complete catalog of *fundamental* truths, to give an *explanatory* account of the world, to tell us how to live our lives, to interpret the world in various ways, to change it). Descriptively, the aim is to say how those goals can be met (through intuitive assessment of thought experiments, by balancing considerations to achieve *epoche,* through endorsing simple theories, by working in the language of the "ontology room"). The excerpts given earlier in this section emphasized the descriptive side of the enterprise; the normative side is more on display in:

[19] Of course, the local criterion may follow from global considerations combined with local facts about the theoretical structure of quantum mechanics.

- The extent to which economics bears on and may be influenced by normative concerns raises **methodological** questions about the relationships between a *positive* science concerning "facts" and a *normative* inquiry into what ought to be. Most economists and methodologists believe that there is a reasonably clear distinction between facts and values, between what is and what ought to be, and they believe that most of economics should be regarded as a positive science that helps policy makers choose means to accomplish their ends, though it does not bear on the choice of ends itself. (Daniel Hausman, *SEP*, "Philosophy of Economics")

The normative side of Theory Selectionism Philosophical Methodology takes center stage in much recent work on ontological commitment (*meta*ontology or *meta*metaphysics), in which a crucial question is what counts as an adequate theory of the world, from which ontological commitments can be read off.

If Epistemologist Philosophical Methodology threatens to relocate Philosophical Methodology entirely within epistemology, to what field of philosophy do methodological questions belong on the current conception? Metaphysics, for example, is responsible for providing a description of the fundamental constituents of the world—but who is responsible for saying what kind of theory metaphysics ought to be providing? Not, on pain of regress, metaphysics itself. Of course, not every philosophical activity needs to have a taxonomic box to contain it. And to some extent Philosophical Methodology, so conceived, will be outsourced to multiple parts of philosophy. (We have already seen the epistemological involvement in the Philosophical Methodology project. Consider also the Williamsonian view that what epistemology should do is to take knowledge as a primitive, rather than seek an analysis of knowledge—this looks to be a piece of Philosophical Methodology driven by considerations drawn from philosophy of language about what constitutes an analysis.) But perhaps there is also a suggestion here that another "meta" category needs to be added to our traditional individuation of philosophical subfields.

6. Hypothesis #5: Necessary Preconditionalism

"Methodological" talk is used quite differently in Giuseppina D'Oro and James Connelly's *SEP* entry on Robin George Collingwood:

- According to Collingwood, neither the proposition "mind exists" nor the proposition "matter exists" is a metaphysical proposition in the traditional sense. They are not metaphysical propositions because they do not assert the existence of metaphysical kinds (mind and matter) but of the **methodological** assumptions that govern the study of mind and nature. These propositions are, as Collingwood puts it, philosophical propositions which define the domains of enquiry or subject matters of the science of history and nature…. As already mentioned, philosophical propositions are not presented as necessary existential claims but as **methodologically** necessary ones. (D'Oro and Connelly, *SEP*, "Robin George Collingwood")

What does it mean to call a proposition "methodologically necessary"? Apparently, that assuming that proposition is a necessary precondition to carrying out theorizing of some form. We see this idea at play, for example, in:

- For logical empiricism, the philosophical significance of relativity theory was above all **methodological**, that conventions must first be laid down in order to express the empirical content of a physical theory. (Ryckman, *SEP*, "Early Philosophical Interpretations of General Relativity")

Roughly, there is *methodological* significance here because there is realization that conventions are a necessary precondition to scientific theorizing, because conventions are needed to mediate between theory and observation in order to have an account of theory confirmation and disconfirmation. This picture of Methodology is explicitly endorsed, for example, in Dever (2006):

- Call Φ a methodological principle for activity A if Φ either is, or is a logical consequence of, a claim whose truth is a constitutive feature of performance of A.[20]

Necessary Preconditionalism provides a satisfyingly "deep" and unified account of what Philosophical Methodology is. Moreover, it's an account that can subsume earlier positions as special cases. It is presumably a necessary precondition on engaging in theorizing of any sort that we have an account of what the conditions of adequacy for the theoretical enterprise are, and of what the epistemic methods are by which justification with respect to the tenets of the theory can be achieved. So Epistemologism and Theory Selectionism can be seen as pursuing particular forms of necessary preconditions.

However, Necessary Preconditionalism is not a satisfying account of Philosophical Methodology without an accompanying account of what exactly necessary preconditions are (or of what constitutive features of activities are), and it is not clear we have such an account. Recall one of our first examples of "methodological" talk—the "sensible methodological adage that 'correlation does not imply causation' " (Stanford, "Underdetermination of Scientific Theory"). In what sense is the claim that correlation does not imply causation a necessary precondition for (one supposes) construction of scientific theories? Since causation is a significant ideological element of many scientific theories, an ability to individuate the causal feature of the world from the merely correlative feature, or the ability to reach justified conclusions about causation on the basis of distributional facts, will certainly be *helpful* in theoretical construction. But is this a *necessary precondition*? Not, it would appear, unless facts about the significant ideological elements of any enterprise are *ipso facto* necessary preconditions for that enterprise. At a minimum, Necessary Preconditionalism seems to require a strong essentialism about philosophical enterprises. If the deployment of causal notions is somehow *essential* to the scientific enterprise, then the "sensible methodological adage" is perhaps plausibly preconditional, but if it is a contingent theoretical discovery that causation is a central notion in scientific theorizing, then preconditionalism is less plausible.

[20] And again, a certain amount of uncomfortable autobiographical reflection is involved here.

7. Hypothesis #6: Hierarchicalism

Consider three more uses of "methodological" talk, each of which has something in common with the "necessary precondition" scheme:

a. There are some complex and rather subtle issues involved in this **methodological** controversy. It is common ground to all these theories, including Hart's, that any attempt to understand what the law is, must rely on a fairly elaborate understanding of law's functions in society, and of the ways in which the law is constituted to fulfill those functions. Furthermore, it seems very plausible to maintain, as Hart himself suggested, that we cannot understand law without understanding the ways in which it is typically regarded by those whose law it is, namely, by those who normally regard the law as giving reasons for their actions. (Andrei Marmor, *SEP*, "The Nature of Law")
b. [Operationalism] is commonly considered a theory of meaning which states that "we mean by any concept nothing more than a set of operations; the concept is synonymous with the corresponding set of operations" (Bridgman 1927, 5). That drastic statement was made in *The Logic of Modern Physics*, published in 1927 by the American physicist P. W. Bridgman…. However, as Bridgman's ideas gained currency they were shaped into a general philosophical doctrine of "operationalism" or "operationism", and in that form became very influential in many areas, especially in **methodological** debates in psychology. (Hasok Chang, *SEP*, "Operationalism")
c. An empirical study in historical science, in the end, cannot do without a metaphysics of history. Bridging irrational reality and rational concept in historical science, or overcoming *hiatus irrationalis* (á la Lask) without recourse to a metaphysics of history still remained a problem as acutely as before. While accepting the broadly neo-Kantian conceptual template as Rickert elaborated it, Weber's **methodological** writings would turn mostly on this issue. (Sung Ho Kim, *SEP*, "Max Weber")

The passage from Marmor appears to trace a methodological issue by holding that claims in philosophy of law depend on prior claims about the function of law in society, and that about the function of law depend on prior claims about reasons for action. Similarly, the Chang passage traces a methodological issue from a starting point in the philosophy of science to the philosophy of language. And Kim moves from a starting point in historical science, to issues about rationality and irrationality, to a "metaphysics of history."

Schematically, all three methodological issues are characterized in terms of dependency hierarchies:

- Law $\Rightarrow_{\text{(depends on)}}$ Collective Action $\Rightarrow_{\text{(depends on)}}$ Rationality
- Science $\Rightarrow_{\text{(depends on)}}$ Language $\Rightarrow_{\text{(depends on)}}$ Epistemology[21]
- History $\Rightarrow_{\text{(depends on)}}$ Epistemology $\Rightarrow_{\text{(depends on)}}$ Metaphysics

[21] The invocation of Epistemology is not overt in the given passage from Chang, but epistemological concerns are a standard point of engagement with operationalist theories of meaning.

This suggests a hierarchicalist picture of Philosophical Methodology: there are chains of dependency relations between different areas of philosophy, and "methodological" considerations are considerations that require retreat from area A up a stage in the hierarchy to area B on which area A depends. The hope here is to put some meat on the Necessary Preconditionalist bones by using the overarching hierarchical structuring of philosophical subdisciplines to determine what the necessary preconditions of any particular theoretical enterprise are. Methodology is, at heart, the right-ordering of the philosophical disciplines.

Hierarchicalism validates what I take to be a common-sense view that methodological concerns are in some way "philosophically deep", because in raising methodological concerns we are (according to Hierarchicalism) moving up the dependency ladder. Unfortunately, the background assumption that there is a suitable partial ordering of the fields of philosophy looks to be wholly without plausibility. Many parts of philosophy have over the years laid claim to the title of "first philosophy", but none of the claimants has made a plausible defense of the title. Which, for example, is prior: metaphysics or epistemology? Are we to believe that epistemology is prior to metaphysics, because we start with how we can know things, and then form a theory of the way the world is that is answerable to the epistemological concerns? Or are we to believe that metaphysics is prior to epistemology, because we start with the nature of the world, and then work out how our epistemic faculties could fit into and engage with that world. Neither picture seems likely to win out over the other, and without a victor, Hierarchicalism is impossible.[22]

8. Interlude: The Puzzle Solved

I have one more hypothesis on the nature of Philosophical Methodology to try out, but before I wrap things up I want to provide a solution to the earlier math puzzle. Order the residents of Springfield in some arbitrary manner. Then each club can be associated with an ordered 1,000-tuple of 0's and 1's—with a position in the tuple filled by 1 if the corresponding Springfield resident is a club member, and by a 0 otherwise. No two clubs can have exactly the same members—the two clubs would both have an odd number of members, and hence an odd number of members in common. Thus 1,000-tuples are uniquely associated with clubs.

Consider these 1,000-tuples as vectors in the vector space on the field {0,1}. Now consider dot products of these vectors. Here are two observations:

a. The dot product of any club vector with itself is 1. The dot product of a vector with itself will produce a sum of 1's for each 1 in the original vector. Since the club must have an odd number of members, this sum must be a sum of an odd number of 1's. But in the field {0,1}, the sum of an odd number of 1's is 1.

[22] There is furthermore a tension between the two thoughts that (a) the "methodological stance" is available in every area of philosophy, and (b) philosophy is, in a roughly Tarskian sense, a semantically closed enterprise. From (a) we will derive that the hierarchy of philosophy is not well-founded, which seems to deprive us of the foundational stance underwriting the universality required by (b).

b. The dot product of any two distinct club vectors is 0. The dot product will produce 1's in its sum only for residents who are members of both clubs. But there must be an even number of such residents. So the dot product is a sum of an even number of 1's. In the field {0,1}, such a sum is 0.

But this shows that the club vectors form part of a basis for the vector space. The vector space is a 1,000-dimensional space, so any basis contains 1,000 vectors. Thus there can be at most 100 club vectors, and hence at most 1,000 clubs. It's then easy to see that this number can be realized.

This is a common pattern in mathematical work. A problem apparently in one area (here, combinatorics) is revealed to be "really" a problem in another area (linear algebra), and a powerful and elegant solution of a previously difficult problem is produced by recharacterizing the problem in the new area and bringing to bear the resources of that other area. It's often a recipe for achieving "deep" insights in mathematics. (Another example: seeing that the compactness of propositional logic can be proved by appeal to Tychonoff's Theorem opens up insights into the relation between logic and topology.) End of interlude (but hold the thought).

9. Hypothesis #7: Eliminatedivism

The final hypothesis about Philosophical Methodology begins with the thought that there's *something* right about Hierarchicalism. "Methodological" or "meta-" is often a label we apply to philosophical moves that invoke the relevance of considerations from philosophical sub-discipline Y to a question in philosophical sub-discipline X. A couple of representative instances:

- Metaethics invokes, roughly, metaphysics and philosophy of language in approaching questions in ethics.
- Meta-ontology invokes, roughly, questions in philosophy of language and epistemology in approaching questions in metaphysics.

It's easy to think we're "moving upward"; getting closer and closer to the deep and fundamental issues.

But consider a sample bit of philosophical reasoning in more detail:

- We're interested in (say) the contextualism/relativism dispute in philosophy of language. We invoke some philosophy of logic to see that the two sides agree on all the moving parts of the logical machinery, on one way of construing it. The disagreement is over which parts are under the control of the semantics, and which under the control of the pragmatics. This leads to the question of where to draw the semantics/pragmatics line, which in turn leads to questions of what aspects of linguistic engagements can be explained through general principles of rational action and which through game-theoretic consequences of information manipulation. In thinking through

> the game theory, we come to realize that the crucial issue is whether information is characterized via a relational or a monadic truth predicate. This transitions to a question in metaphysics: how substantive is our notion of truth, and what features of the world does it answer to? Examining a substantive notion of truth leads to developing a theory of facts, which in turn leads to consideration of the Bradley regress argument. We're then put back into semantics, as we consider the role of unsaturated senses in resolving this argument. And so on.

There's no *hierarchy* here. We move from philosophy of language to philosophy of logic to metaphysics, but then from metaphysics back to philosophy of language. So there is movement from one area of philosophy to another, and movement that is, in some important sense, *methodologically driven*. But the movement isn't the tracing of a pre-given hierarchical structuring of the field. It is rather, as in the example of The Puzzle, part of the ongoing attempt to do better work, and achieve more and deeper insights, by thinking about how various components of the field relate to one another, rather than allowing one's work to reside entirely in one subdiscipline of philosophy.

Here, then, is a final picture. Philosophical Methodology is the study of philosophical method: how to do philosophy well. But at the end of the day there isn't much to say about how to do philosophy well. There are some general guiding aphorisms, and "be on the lookout for the relevance of epistemology to your questions in philosophy of language" is one of them. But the guiding aphorisms guide only very loosely—there's no recipe telling you precisely when to invoke epistemological considerations in linguistic theorizing. There's only the skillful receptiveness to possible fruitful interactions. On this view, there's a half-triumph for the first hypothesis of Eliminativism. There *is* such a thing as Philosophical Methodology—it's the study of the rules of good philosophical practice. But there are no rules of good philosophical practice, so we have instead Eliminativism about philosophical methodology. Philosophical Methodology is not itself eliminated, but has an eliminated topic.

APPENDIX

Entries in the *Stanford Encyclopedia of Philosophy* That Use the word "Methodological"

Methodological Individualism, Naturalism, Naturalism in Philosophy of Mathematics, Philosophy of Economics, Behaviorism, Physicalism, Historicist Theories of Rationality, Reductionism in Biology, Robin George Collingwood, Max Weber, Unity of Science, Holism and Nonseparability in Physics, Formal Learning Theory, Creationism, The Incommensurability of Scientific Theories, The Pure Theory of Law, The Nature of Law, Early Philosophical Interpretations of General Relativity, Naturalism in Legal Philosophy, Simplicity, Dedekind's Contributions to the Foundations of Mathematics, Thomas Kuhn, Analysis, Collective Responsibility, Philosophy of Biology, Feminist History of Philosophy, Feminist Perspectives on Science, Lvov-Warsaw School, Identity and Individuation in Quantum Theory, Quantum Gravity, Paul Feyerabend, Analytic

Feminism, Comparative Philosophy: Chinese and Western, Hermeneutics, Hume's Naturalism and Anti-Naturalism, Animal Consciousness, Empathy, Theory and Bioethics, Cosmology: Methodological Debates in the 1930s and 1940s, Max Horkheimer, Mathematical Style, Moral Non-Naturalism, Multiple Realizability, Scientific Progress, Symmetry and Symmetry Breaking, Copenhagen Interpretation of Quantum Mechanics, Religion and Science, Cognitive Science, The Kokugaku, The Theology of Aristotle, John Stuart Mill, Aristotle's Biology, The Biological Notion of Self and Non-Self, Darwinism, Discourse Representation Theory, Edmund Husserl, Monism, Operationalism, Underdetermination of Scientific Theory, Scientific Realism, Philosophy of Technology, Teleological Theories of Mental Content, Personal Identity and Ethics, Moral Psychology: Empirical Approaches, Bell's Theorem, Rene Descartes, Jurgen Habermas, The Philosophy of Neuroscience, Evolution, Feminist Epistemology and Philosophy of Science, Colonialism, Naturalized Epistemology, Business Ethics, Pragmatist Feminism, Hans-Georg Gadamer, Charles Hartshorne, James Mill, Karl Jaspers, Karl Marx, Richard Rorty, Scientific Revolutions, Aristotelianism in the Renaissance, Approaches to Feminism, Implicature, Public Justification, Joseph Kaspi, Communitarianism, Externalism About Mental Content, Dewey's Moral Philosophy, Ecology, Philosophy of Education, Einstein's Philosophy of Science, Francis Bacon, Distributive Justice, The Analysis of Knowledge, Combining Logics, Paul Natorp, Plato, Alfred Schulz, Thermodynamic Asymmetries in Time, Wilhelm von Humboldt, Vienna Circle, Logic and Artificial Intelligence, Arabic and Islamic Philosophy of Language and Logic, Attention, Russell's Logical Atomism, Moral Epistemology, Experiment in Physics, Units and Levels of Selection, Albert the Great, Aristotle, Aristotle on Causality, Luitzen Egbertus Jan Brouwer, Critical Theory, Evolutionary Epistemology, The Frame Problem, Pierre Gassendi, Johan Gottfriend von Herder, Hobbes's Moral and Political Philosophy, Epistemic Logic, Indispensibility Arguments in the Philosophy of Mathematics, Abilities, Aristotle's Psychology, Auguste Comte, Etienne Bonnot de Condillac, Pierre Duhem, John Duns Scotus, Ernst Mach, Feminist Social Epistemology, Gadamer's Aesthetics, Herman von Helmholtz, Hilbert's Program, Japanese Zen Buddhist Philosophy, Kant's Philosophy of Science, Law and Language, Legal Obligation and Authority, Models in Science, Nothingness, William of Ockham, Panpsychism, Personalism, Phenomenology, Pseudo-Dionysius the Areopagite, Carl Stumpf, Truthlikeness, 18th Century French Aesthetics, Joseph Albo, Antonio Rosmini, Peirre Bayle, Philosophy for Children, Contradiction, Decision-Making Capacity, Defaults in Semantics and Pragmatics, Doing vs. Allowing Harm, Epiphenomenalism, Shem Tov Ibn Falaquera, Feminist Metaphysics, Michel Foucault, Galileo Galilei, Ibn Kammuna, The Historical Controversies Surrounding Innateness, William James, The Economic Analysis of Law, The Natural Law Tradition in Ethics, Simpson's Paradox, Francesco Patrizi, John Philoponus, Giovannia Pico della Mirandola, Platonism in the Philosophy of Mathematics, Science and Pseudo-Science, Reid on Memory and Personal Identity, Saadya, The Social Dimensions of Scientific Knowledge, Scottish Philosophy in the 19th Century, Wilfrid Sellars, Chauncey Wright, Bolzano's Logic, Roderick Chisholm, Feminist Philosophy of Religion, Intentionality, Leibniz's Philosophy of Physics, Stanislaw Lesniewski, Philosophy of Mathematics, Quantum Field Theory, Roger Bacon, Eugen Rosenstock-Huessy, Franz Rosenzweig, The Problem of Induction, Neutral Monism, Relativism, Animal Cognition, Descriptions, Plato's Ethics: An Overview, Henry Sedgwick, Epistemological Problems of Testimony, Cosmological Argument, Nicholas of Cusa, Delusion, Descartes' Physics, Dialetheism, Wilhelm Dilthey, Feminist Perspectives on Disability, Feminist Bioethics, Feminist Perspectives on Power, Fictionalism, Herman Lotze, Identity, The Distinction Between Innate and Acquired Characteristics, Liberalism, Locke's

Philosophy of Science, The Logic of Action, Mental Illness, Maurice Merleau-Ponty, Moral Relativism, Is Either Moral Realism or Moral Anti-Realism More Intuitive Than the Other, Natural Law Theories, Postmodernism, Pragmatism, Private Language, Petrus Ramus, Paul Ricoeur, Gilbert Ryle, August Wilhelm von Schlegel, Phenomenological Approaches to Self-Consciousness, Naturalistic Approaches to Social Construction, Space and Time: Inertial Frames, Species, Alfred Tarski, Teleological Arguments for God's Existence, Wihelm Maximilian Wundt, Bernard Bolzano, Introspection, The Kyoto School, Mental Imagery, Saint Thomas Aquinas, Aquinas' Moral, Political, and Legal Philosophy, Isaiah Berlin, The Metaphysics of Causation, Probabilistic Causation, Epistemic Contextualism, Evidence, Kant's Philosophical Development, Kant and Hume on Causality, Alexius Meinong, Molecular Genetics, Moral Anti-Realism, Auditory Perception, Method and Metaphysics in Plato's Sophist and Statesman, Pornography and Censorship, John Rawls, Social Institutions, Structural Realism, Temporal Parts, Truth Values, Bernard Williams, Game Theory, Aristotle's Ethics, Chaos, Consciousness, Continuity and Infinitesimals, Analytic Philosophy in Early Modern India, Kurt Godel, Thomas Hill Green, International Justice, Independence Friendly Logic, Molecular Biology, Newton's Philosophiae Naturalis Principia Mathematica, Pain, Paradoxes and Contemporary Logic, Philosophy of Religion, Scientific Explanation, Sidney Hook, Absolute and Relational Theories of Space and Motion, Empathy: The Study of Cognitive Empathy and Empathic Accuracy, Innateness and Language, Mental Imagery: European Responses of Jaensch, Freud, and Gestalt Psychology, Mental Imagery: Founders of Experimental Psychology Wilhelm Wundt and William James, Experiment in Physics Appendix 4: The Fall of the Fifth Force, Thomas Reid, 18th Century German Aesthetics, Pleasure, David Lewis's Metaphysics: The Contingency of Humean Supervenience, Sociobiology: Construction of Sociobiological Explanations, Analysis: Conceptions of Analysis in Analytic Philosophy, Analysis: Definitions and Descriptions of Analysis

To give a rough classification of these articles: 97 are in history of philosophy (with all articles devoted to specific persons counted as history), 53 are in philosophy of science (20 in general philosophy of science, 19 in philosophy of physics, and 14 in philosophy of biology), 28 are in philosophy of mind, 15 are in ethics, 14 are in epistemology, 13 are in metaphysics, 12 are in philosophy of language, 12 are in feminist philosophy, nine are in logic, nine are in philosophy of mathematics, seven are in non-Western philosophy, six are in philosophy of law, five are in political philosophy, two are in philosophy of action, two are in philosophy of education, and one is in Continental philosophy. One interesting question (which I won't attempt to answer here) is why "methodological" talk is so heavily concentrated in historical pieces. The phenomenon is partially explained by the prevalence of named-figure articles in the SEP devoted to figures from the philosophy of science (the strong representation of philosophy of science in "methodological" talk was addressed in section 5), but that partial explanation leaves a substantial phenomenon unexplained.

References

Allen, Colin, "Animal Consciousness", *The Stanford Encyclopedia of Philosophy* (Winter 2011 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2011/entries/consciousness-animal/>, accessed September 9, 2015.

Baker, Alan, "Simplicity", *The Stanford Encyclopedia of Philosophy* (Fall 2013 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2013/entries/simplicity/>, accessed September 9, 2015.

Bickle, John, "Multiple Realizability", *The Stanford Encyclopedia of Philosophy* (Spring 2013 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2013/entries/multiple-realizability/>, accessed September 9, 2015.

Brigandt, Ingo and Love, Alan, "Reductionism in Biology", *The Stanford Encyclopedia of Philosophy* (Summer 2012 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2012/entries/reduction-biology/>, accessed September 9, 2015.

Cassirer, Ernst (1921), *Einstein's Theory of Relativity*. Chicago: Open Court.

Cat, Jordi, "The Unity of Science", *The Stanford Encyclopedia of Philosophy* (Summer 2013 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2013/entries/scientific-unity/>, accessed September 9, 2015.

Chang, Hasok, "Operationalism", *The Stanford Encyclopedia of Philosophy* (Fall 2009 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2009/entries/operationalism/>, accessed September 9, 2015.

Dever, Josh (2006). "Compositionality". In Ernest Lepore and Barry Smith, eds., *The Oxford Handbook of Philosophy of Language*, pp. 633–666. Oxford University Press.

D'Oro, Giuseppina and Connelly, James, "Robin George Collingwood", *The Stanford Encyclopedia of Philosophy* (Summer 2010 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2010/entries/collingwood/>, accessed September 9, 2015.

Faye, Jan, "Copenhagen Interpretation of Quantum Mechanics", *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2008/entries/qm-copenhagen/>, accessed September 9, 2015.

Finnis, John, "Natural Law Theories", *The Stanford Encyclopedia of Philosophy* (Fall 2011 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2011/entries/natural-law-theories/>, accessed September 9, 2015.

Geurts, Bart and Beaver, David I., "Discourse Representation Theory", *The Stanford Encyclopedia of Philosophy* (Fall 2011 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2011/entries/discourse-representation-theory/>, accessed September 9, 2015.

Hausman, Daniel M., "Philosophy of Economics", *The Stanford Encyclopedia of Philosophy* (Spring 2013 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2013/entries/economics/>, accessed September 9, 2015.

Kim, Sung Ho, "Max Weber", *The Stanford Encyclopedia of Philosophy* (Fall 2012 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2012/entries/weber/>, accessed September 9, 2015.

Marmor, Andrei, "The Nature of Law", *The Stanford Encyclopedia of Philosophy* (Winter 2011 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2011/entries/lawphil-nature/>, accessed September 9, 2015.

Ramberg, Bjørn and Gjesdal, Kristin, "Hermeneutics", *The Stanford Encyclopedia of Philosophy* (Summer 2013 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2013/entries/hermeneutics/>, accessed September 9, 2015.

Russell, Bertrand (1918), *The Philosophy of Logical Atomism*. New York: Routledge.

Ryckman, Thomas A. (2005), *The Reign of Relativity: Philosophy in Physics 1915-1925*. Oxford: Oxford University Press.

Ryckman, Thomas A., "Early Philosophical Interpretations of General Relativity", *The Stanford Encyclopedia of Philosophy* (Fall 2012 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2012/entries/genrel-early/>, accessed September 9, 2015.

Shields, Christopher, "Aristotle's Psychology", *The Stanford Encyclopedia of Philosophy* (Spring 2011 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2011/entries/aristotle-psychology/>, accessed September 9, 2015.

Sorensen, Roy, "Nothingness", *The Stanford Encyclopedia of Philosophy* (Winter 2012 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2012/entries/nothingness/>, accessed September 9, 2015.

Stanford, Kyle, "Underdetermination of Scientific Theory", *The Stanford Encyclopedia of Philosophy* (Winter 2013 Edition), Edward N. Zalta (ed.), forthcoming URL = <http://plato.stanford.edu/archives/win2013/entries/scientific-underdetermination/>, accessed September 9, 2015.

Stebbing, Susan (1932). "The Method of Analysis in Metaphysics". Proceedings of the Aristotelian Society 33: 65–94.

Stueber, Karsten, "Empathy", *The Stanford Encyclopedia of Philosophy* (Summer 2013 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2013/entries/empathy/>, accessed September 9, 2015.

Szabó, Zoltán Gendler, "Compositionality", *The Stanford Encyclopedia of Philosophy* (Fall 2013 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2013/entries/compositionality/>, accessed September 9, 2015.

Thagard, Paul, "Cognitive Science", *The Stanford Encyclopedia of Philosophy* (Fall 2012 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2012/entries/cognitive-science/>, accessed September 9, 2015.

Tuana, Nancy, "Approaches to Feminism", *The Stanford Encyclopedia of Philosophy* (Spring 2011 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2011/entries/feminism-approaches/>, accessed September 9, 2015.

PART II

TRADITIONS AND APPROACHES

CHAPTER 2

THE METHODOLOGY OF THE HISTORY OF PHILOSOPHY

CALVIN G. NORMORE

1. Introduction

What constitutes the History of Philosophy is as contested a territory as what constitutes Philosophy itself.[1][2] This is, in part, because each is used to legitimate approaches to the other, and to such other disciplines as Politics and Theology.[3] Indeed in certain quarters, most famously Hegelian, Philosophy and the History of Philosophy are so intertwined that to contest the one just is to contest the other. This contested state makes it nigh impossible to do justice in a short compass to the different understandings and approaches to the History of Philosophy now current so, heart on sleeve, I offer in this chapter what might be called a strong reading both of how things are and of how they should be and a (very limited) defence of that reading.

First then the issues: one is whether the History of Philosophy (done well let us suppose) is Philosophy; another is whether it is History. An issue related to both is whether the history of Philosophy can be studied 'internally' or whether it is best studied as (say) Intellectual or Cultural History with a particular focus. At issue here, of course, is the importance of what has come to be called 'context' in the study of the history of Philosophy. I shall argue for aiming to present the history of Philosophy internally as a default.

My reading has consequences. Perhaps the most striking is that what passes for the History of Philosophy in most quarters with which I am familiar (and indeed most of my

[1] Many of the ideas in this chapter were developed in a series of seminars jointly taught with Brian Copenhaver over the past decade. and both the chapter and I owe him a special debt. Approaching these issues as a trained historian, Brian has a different perspective on them and remains, I think, sceptical that an 'internal' history of philosophy is possible. We will see!

[2] Calvin G. Normore is Professor of Philosphy, UCLA, Emeritus Macdonald Professor of Moral Philosophy, McGill University, and Honorary Research Professor, the University of Queensland.

[3] For a little on this i*deological* use of the History of Philosophy, see C. G. Normore, 'Doxology and the History of Philosophy', *Canadian Journal of Philosophy* Supplementary volume 16, 1990 pp. 203–26.

own study of the past of Philosophy) is not quite History of Philosophy at all. Nor, I argue, should it be. Nonetheless there may be a place for what really is History of Philosophy, and confusing it with its near kin just breeds confusion.

2. THEORY

2.1.1 What Then is the History of Philosophy?

Philosophy has a considerable past. There are many ways of studying that past.[4] I understand the subject which is The History of Philosophy to study that past in one family of ways.[5] I take it there are other subjects, for example Intellectual History and Social History, which study it in other ways. I take it too that there is something, not quite a subject in its own right, which I have called Doxology, and something else, also not quite a subject, that I will call Anthropology of Philosophy, which study items in the past of Philosophy in ways useful to, but not identical with, the History of Philosophy. Because both what counts as Philosophy and what counts as History of Philosophy are contested, many of those who practice Intellectual or Social History or Doxology or Anthropology of Philosophy insist that they are Historians of Philosophy. That they study the past of Philosophy I've no wish to deny, but I want also to insist that there is (and perhaps, just perhaps, should be) a distinct subject with a distinct history and a distinct set of concerns which has a legitimate claim to be History of Philosophy, and that it is what is in fact presuppposed in current historiography of the History of Philosophy.[6]

[4] Here I am heavily indebted to unpublished lectures of Michael Frede delivered at UC Riverside in 1991 and to the Introduction to M. Frede *Essays in Ancient Philosophy*, University of Minnesota Press, Minneapols, 1987.

[5] A terminological convention. In general I distinguish a subject from its subject matter by capitalizing the name of the subject and (except when it begins a sentence) not that of its subject-matter. Thus Mathematics with a majuscule 'M' is the subject which studies mathematics with a miniscule 'm'. The subject, Philosophy, is a special case because, inter alia, it studies itself, thus providing fertile ground for confusion. Still, Philosophy studies what makes up various philosophies which taken together constitute philosophy. The History of Philosophy is a subject whose subject matter is the history of Philosophy. If what makes up philosophies were to have a history of their own, then there could be a subject, the History of philosophy, which studied their histories but, as I see it, while which items Philosophers (those who practise Philosophy) count as philosophy changes over time, those items themselves do not. As I understand it, whatever is expressed now by the opening sentence of Aristotle's *Categories* is what was expressed by it when Aristotle penned it, whatever we or he may wish to or be able to make of it. As for the history of philosophy (all uncapitalized), I fully agree with Rorty, Schneewind, and Skinner when they write 'There is nothing general to be said in answer to the question "How should the history of philosophy be written?" except "As selfconsciously as one can—in as full awareness as possible of the variety of contemporary concerns to which a past figure may be relevant" ' (R. Rorty, J. B. Schneewind, and Q. Skinner, *Philosophy in History* (Cambridge University Press, Cambridge, UK, 1984), p. 11.

[6] This is (of course!) itself controversial. Someone might think that the meaning of the first line of the *Categories*, even as Aristotle penned it, depends on the context in which it is embedded and so is determined by the history of that text. I'm not one of them. This is also not to deny that *assertions* of bits of philosophy have a history. Susan Neiman reminds us that 'One thing uniting new historians is the reminder that earlier philosophers wrote in response to events in real time' (Susan Neiman,

Despite the disagreement about what it is, there is widespread agreement among those who study the history of the History of Philosophy that it is a distinct subject and a relatively recent phenomenon. Some, like Emile Bréhier or Conal Condren, date its beginnings to the critical turn taken by European scholars of religion in the seventeenth century.[7] Others, like Michael Frede, date it to the efforts of eighteenth-century scholars like Fullerborn. The monumental *Storia delle storie generali della filosofia* details discussions of the past of Philosophy beginning with antiquity, but it too recognizes a fundamental shift somewhere in the Early Modern period. The editor of the English translation of Volume II puts it this way:

> The volume which we present here in its English translation . . . concerns a particularly significant (we could almost say 'strategic') phase in the development of modern philosophical historiography, which in the period between the second half of the seventeenth and the first half of the eighteenth century (from Descartes to Brucker, precisely) abandoned its philological and erudite guise and took on the form of a 'critical' and 'philosophical' history of philosophy, in a complex and problematic interchange with the concerns of modern philosophy (represented in particular by Descartes, Leibniz, and Locke), but also with the nascent histoire de l'esprit humain. Leaving aside the play on words suggested by formulas such as the 'philosophical history of philosophy' or the 'philosophy of the history of philosophy', we see a true change in intentions and methods which was fundamentally to influence modern cultural sensitivity and was to develop finally into the Hegelian apotheosis of the unity of philosophy and history of philosophy, but also, in another sense, into the methodology of 'intellectual history'.[8]

This relatively recent dating and the disagreements about which recent dates are the relevant ones seem connected with the view that, while Philosophy is ancient, the discipline we now call History is a recent development, and with the expectation that anything properly called History of Philosophy will be History as well as Philosophy.

This Janus-faced character is distinctive of the History of Philosophy. Few expect the History of Chemistry to be of professional interest to contemporary chemists but Historians of Philosophy are typically housed with Philosophers and are expected to interact with Philosophers on their own terms.

Upon a little reflection this should not be surprising. Though there are significant exceptions, most of those who become interested in the past of Philosophy do so by first taking up Philosophy, then being introduced to texts from the past of Philosophy, then finding those texts fascinating and puzzling and setting out to understand them. How the texts fit

'Meaning and Metaphysics' in Jerome B. Schneewind *Teaching New Histories of Philosophy* (Princeton University: Princeton, 2004), p. 40).

Some did but some did not! If Aristotle or Plotinus were responding to contemporary events in real time they hid it well!

[7] Cf. Emile Bréhier, *Histoire de la Philosohie* t. 1 (Librairie Félix Alcan, Paris, 1928), Introduction III, pp. 17ff. (There is an excellent electronic edition prepared by Peirre Palpant available freely at <http://classiques.uqac.ca/classiques/brehier_emile/Histoire_de_philo_t1/brephi_1.pdf>, accessed September 12, 2015. Cf. also Conal Condren, *Hobbes, the Scribblerians and the History of Philosophy* (Pickering and Chatto, London, 2012), especially the Introduction and Afterword and the literature referenced therein.

[8] *Models of the History of Philosophy Volume II: From the Cartesian Age to Brucker* ed. Gregorio Piaia and Giovanni Santinello (Springer, Dordrecht, 2011).

together to form elements of a history and how to understand that history, comes later, if at all, and flows from efforts to understand particular texts.

Why might anyone want more than such understanding of philosophical texts? Historically, the answer to that question seems closely connected with the contest over the nature of Philosophy itself. From Brucker's argument that Philosophy began to make progress only when, with (Francis) Bacon and his successors, it became eclectic, to Heidegger's 'history of Being' and beyond, histories of Philosophy have almost universally been efforts to argue for particular ways of doing Philosophy itself. Such considerations might suggest that the History of Philosophy is best ignored by those more or less satisfied with the conception of Philosophy in which they are institutionally immersed and that they (we?) would be well advised to stick to Doxology and Anthropology of Philosophy. I want to suggest, however, that such reaction is over-reaction and that there is indeed a place among Philosophers as well as among Historians for History of Philosophy as its current historiographers understand it. First though we need some background.

Philosophy is a *discipline*. To practice it is, willy-nilly, to belong to a community of practitioners. To count someone as a philosopher is to count them into such a community. Though such a community is bound together by a common focus on philosophies, there is much more to it than that focus. Philosophers have often shared a distinctive institutional life and sometimes even a distinctive mode of dress![9]

The discipine of Philosophy has a history. Moreover, the past of Philosophy—of the set of practices, people, events, concepts, theses, and arguments which make up that discipline—is not the past of the collection of philosophies, which is both narrower and wider. On the one hand, a philosophy does not include (usually anyway) the practices, people, and events which are included in Philosophy but only the concepts and theses and the arguments and other ways of presenting considerations for and against them. Sociologists may attempt to discover whether philosophers differ from others in their ethical views; philosophers qua philosophers typically do not. On the other hand, a philosophy may include many views and arguments and other considerations which were part of Philosophy but are no longer, and very likely many which are not yet part of Philosophy but may become so.[10] While it seems that philosophies do not have a past any more than numbers do, what *counts* as a philosophy has a past, just as what counts as a part of mathematics does, and that is, I think,

[9] The view that to understand the Philosophy of antiquity one must see it as a way of life has been defended, especially by Pierre Hadot. That to understand the Philosophy of the early modern era one has to see it in terms of the persona of the philosopher has been argued by Conal Condren, Stephen Gaukroger, and Ian Hunter. For Hadot's views, cf. Pierre Hadot, *Philosophy as a Way of Life* (Blackwell, Oxford, 1995), and *What is Ancient Philosophy?* (Harvard University Press, Cambridge, MA, 2002). For Condren, Gaukroger, and Hunter, cf. C. Condren, S. Gaukroger, and I. Hunter (eds), *The Philosopher in Early Modern Europe: The Nature of a Contested Identity* (Cambridge University Press, Cambridge, 2006).

[10] The view that there is a *dator formarum* (perhaps related to the moon), which accounts *inter alia* for the generation of maggots, used to be part of Philosophy but is no longer. Someone who maintained it now would almost certainly not thereby be a philosopher but a crackpot. Still, there are significant and fairly recent philosophies, Francesco Suarez's for example, of which it is a part. Whether Freud's account of the Oedipal complex is part of Philosophy is contested, and so is whether Freud should be counted a philosopher.

because Philosophy is what Ian Hacking has dubbed a human kind—a kind shaped in part by our conception of it.[11] As Philosophy changes, what we count as philosophy also does.

Because what counts as Philosophy has varied (sometimes considerably) over time, what it is to study the past of Philosophy is itself open to determination. We can be interested in what of the past we would here and now count as philosophy or in what would then have counted as philosophy. The results may be very different. Contemporary English translations of Descartes' *Principles of Philosophy* typically leave out much of what we woud now call Descartes' Physics. These parts of the texts are no longer thought to be of Philosophical interest even though these were closely studied by such writers as Spinoza and Newton. Whether they will be included in a story of the past of Philosophy will depend on the orientation of the storyteller. As I see it, Historians of Philosophy can go either way—though perhaps not taking their audience with them!

How much latitude is there in what might be counted as Philosophy? One could of course just count what contemporary Philosophers have already found interesting—the 'canon' or something like it. Another way of fixing what is to be counted would be to track the history of the use of the word 'philosophy', its cognates, and words explicitly introduced to translate it. Interesting though that project might be, it would be unlikely to come close to tracking either what is currently included in histories of Philosophy or what those seeking to reform what is included would wish to see there. Yet another approach would be to start from what is here and now considered philosophy and to work back, taking into account the past figures we take to be philosophers, the figures they took to be philosophers, the figures they in turn took to be philosophers, and so on.[12] No doubt creative Historians of Philosophy will find yet other approaches.

Sometimes at least, philosophers hold the views they proffer on account of what they or we would regard as philosophical considerations or *reasons*. What counts as a philosophical consideration or reason is contested territory (of course?) but at least within subcommunities of Philosophers there is usually some consensus. Philosophers sometimes advance views not in their own voices but to see what merit may be found in them and/or what considerations count for and against them but, again, usually the views philosophers develop over time and at some length tend to be those they proffer in their own voices. If philosophers do hold the views they proffer (or even if we are entitled to suppose they do), and if belief aims at the truth and if truth requires coherence, we may suppose that philosophers develop and change their views in the light of arguments and other considerations, aiming at such philosophical goals as coherence and truth. If this is so, then to the extent that one does not understand and appropriately weigh these considerations and arguments or does not appreciate the role of these goals, one will not understand why the past of Philosophy is as it is. On the other hand, if to understand and weigh those considerations requires understanding how they interact with other considerations of which the author being studied is aware but which are not explicit in the texts studied, then to the extent that one does not understand the Philosophical milieu in which an author is working, one will

[11] See for example, Ian Hacking, 'The Looping Effects of Human Kinds', in *Causal Cognition,* ed. D. Sperber, D. Premack, and A. James Premack (Oxford University Press, New York, 1996), pp. 351–83.

[12] This approach would get Mencius into the tent without including (all of?) the authors of the self-help books that populated the Philosophy sections of bookstore chains—when there were bookstore chains!

not understand what is being studied. There is, then, a case to be made for a subject which attends to the arguments and other considerations explicit in texts in the context of considerations which appear in historically related texts, in the context of considerations about what best explains why a text is at it is given that the author(s) is/are aiming at such philosophical goods as coherence and truth.

That subject, say I, is the History of Philosophy. Since the Historian of Philosophy can take as Philosophy either what we now would or what was once so taken, it is in part an empirical matter how wide a spectrum of approaches the History of Philosophy will support. If what counts as Philosophy or what counts as a philosophical consideration or reason varies widely from time to time so may Histories of Philosophy.

Given this latitude about how to proceed, a central methodological issue which must be faced by any Historian of Philosophy is what questions to ask. It is a commonplace that to understand a text (or a practice) one should approach it with questions—but which questions? As just suggested, at least two options present themselves: 1) ask what you would ask of a text by one of your philosopher contemporaries, or 2) ask what one of the writer's philosopher contemporaries would (as far as you can tell) have asked.

Both of these approaches have respectable followings. Here, for example, is the historian of Ancient Philosophy Paolo Crivelli meditating on his own approach to Aristotle's text in his *Aristotle On Truth*:

> In the case at hand, my approach involves engaging in a philosophical discussion with Aristotle. I ask him some of the questions about truth which many modern analytic philosophers are interested in. One question I ask Aristotle is: 'What are the bearers of truth or falsehood?' Another question is: 'What are the truth conditions for predicative assertions?' Sometimes I ask Aristotle a further question raised by his answer to one of these questions. One of these further questions is: 'How can your truth conditions for predicative assertions deal with what seem predicative assertions with an 'empty' subject or predicate?'
>
> It should be clear how this approach will contribute to achieving one of this book's primary aims, the aim of gaining a philosophical understanding of Aristotle's views on truth: asking Aristotle some of the questions about truth which many modern analytic philosophers are interested in is evidently a way to gain a philosophical understanding of Aristotle's views on truth. But it is not obvious that this approach should contribute to the other primary aim of the book, the reconstruction of all of Aristotle's most significant views on truth and falsehood: one can imagine situations where by asking a philosopher 'analytic' questions about truth one would completely fail to reconstruct that philosopher's most significant views on the matter. However, as it happens, asking 'analytic' questions bears fruit in Aristotle's case: it does allow me to reconstruct almost all of his most significant views on truth and falsehood.
>
> (Crivelli 2004, p. 39)[13]

[13] Closer to the centre of the subject as it is now practised is Michael Frede's essay, 'The Original Notion of Cause'. Frede begins the essay with reflection, not on questions a contemporary philosopher as such might ask, but on a puzzle a contemporary philosopher might raise about the history of philosophy:

> However muddled our notion of a cause may be it is clear that we would have difficulties in using the term 'cause' for the kinds of things Aristotle calls 'causes'.1 We might even find it misleading to talk of Aristotelian causes and wonder whether in translating the relevant passages in Aristotle we should not avoid the term 'cause' altogether. For an end, a form, or matter do not seem to be the right kinds of items to cause anything, let alone to be causes.
>
> (Frede 1987, p. 125)

Crivelli here puts nicely two aims of the History of Philosophy 1) to 'gain a philosophical understanding' of a body of text and 2) to reconstruct all the most significant views of a philosopher on a given subject. Given those aims, Historians of Philosophy must seek to make the texts which they study intelligible to *themselves* and *their* audience. Historians of Philosophy can accept an interpretation of a text which has it not make sense *to themselves* only if they have an explanation of why the text might not make sense—say because it is corrupt or because the writer deliberately or inadvertently made it incoherent. In any of those cases the Historian of Philosophy studies that text only to access another implicit text which the author really wrote or really would have written were it not for extra-philosophical factors. Moreover, if there is an interpretation which makes sense of the text, a Historian of Philosophy must prefer it to a search for an implicit text behind the text at hand. Historians of Philosophy must also be prepared to test any interpretation they do propose against *all* the relevant text. An interpretation that is incompatible with some of the relevant text is thereby disqualified as an interpretation of any of it.

These may sound like platitudes but they bear directly on current practice among those studying the history of Philosophy. The first aim is incompatible with at least one understanding of the approach of Leo Strauss and his followers, and the second with at least one strand of postmodern literary theory.

Making sense of a text in terms we can understand (which need not of course mean making sense of it in terms that can be understood without studying it in its historical context) can be a complex and difficult process. What, for example, is meant by Aristotle's claim that in thought the intellect becomes what is thought about? Both Aristotle's Greek and our English is being used here in unfamiliar ways and it is a task of the Historian of Philosophy to explain these uses to *us*. Merely repeating or using a dictionary translation of Aristotle's Greek won't do. Such explanation may take the historian far afield—into related uses of the Greek, related earlier and later texts, and excursions into contemporary Philosophy of Mind. We may come away still with the sense that there is more here than yet understood, but unless we understand the claim to some extent, the Historian of Philosophy has left the job undone.

Making sense of a text in terms *we* can understand is a necessity in the History of Philosophy but it is not the whole story. Philosophers in the past asked questions we no longer ask (and perhaps cannot reasonably ask any longer). Philosophers in the past also did not ask questions we find completely natural.[14] Understanding why the questions asked

He turns, however, at once to a meta-question about the source of the puzzle, and this leads him to reflect on the historical development of the Greek terms *aition* and *aitia* and their translations as *causa* and *cause*. To trace this development he relies not only on Philosophical texts but on material from the history of medicine (which never was part of Philosophy) and the history of Physics (which once was but seems to be no longer). He assembles these to argue that it was the Stoic concern with attributing responsibility and with factoring out the different senses in which contributors to an event might be responsible for it which led to our conception of causing as making something happen. All of that is respectable (if sometimes quite speculative) History. Nonetheless, Frede presents this History as helping us understand the notion of cause we currently have (by explaining how *it* came about).

[14] Many recent discussions of knowledge (Bayesian and 'formal') rely heavily on the concept of probability. Richard Jeffrey once offered a prize to anyone who could find a discussion of probabilities before the sixteenth century. To the best of my knowledge it is unclaimed.

were asked and those not were not, and understanding what was and would have been taken as satisfactory answers is also central to the History of Philosophy as I have demarcated it.

It is here that Historical sensitivity and historical knowledge play an especially important role. When Descartes proposes to Princess Elizabeth that she look to scholastic accounts of gravitas for an understanding of how the mind moves the body he is alluding to a long tradition.[15] To understand the notion of *gravitas* as he did we may need to canvas much of that tradition. We may need to study, in some detail, a considerable body of medieval Physics.

This project is, however, a very different one from that which underlies much historical work about Philosophy. Such historical work does not aim to explain such things as Descartes' conception of *gravitas*, but aims instead to explain why Elizabeth thought to ask how the mind moves the body, and why Descartes thought to refer her to the scholastic tradition as if it would help. These are interesting questions, and someone who can answer them has a deeper understanding of the past of Philosophy than someone who cannot, but they are not *philosophical* questions and they would, in the terminology I propose, be part of Intellectual History not the History of Philosophy.

The History of Philosophy is one approach to the study of the past of Philosophy. It is not the only subject studying the past of Philosophy, nor do I see any reason to think it is the only way it should be. Nonetheless, among Historical approaches to the past of Philosophy it does (and I shall argue should) hold a central place.

That it does is an institutional fact. The past of Philosophy is, as a matter of fact, studied chiefly in Philosophy Departments, schools, and centres.[16] Its study in Philosophy departments, schools, and centres is shaped to a considerable extent by the interests of the philosophers and students of philosophy who make those up. Many of these philosophers and students are professionally interested in the history of Philosophy just insofar as it connects or could come to connect with contemporary Philosophy. Thus any approach to the past of Philosophy which can be embedded in existing institutional structures will have to be one which connects with contemporary philosophical concerns. What must such an approach be like?

Philosophy thrives dialectically and philosophers are interested in the views of other philosophers. The study of such views in the way philosophers typically study the views of other philosophers is what I call Doxology.[17] Doxologists have an interest in attributing to the philosophers they study the views they actually hold (though that may not always be their most salient interest), and to the extent that this requires understanding the contexts in which they are formulated, doxologists have an interest in understanding those contexts. If the philosopher whose view is being studied is dead, then understanding the context in which that philosopher worked may require knowing some history and so may require some Historical sophistication. Nonetheless, while one may need to be a Historian to understand why Aristotle felt a need to construct a modal syllogistic, one need not be a Historian to understand (say) the structure of that syllogistic itself.

[15] There are twentieth-century pictures related to it (I think here of the work on G. H. von Wright on action) but much will be missed if one's focus is on them!

[16] Was it always so?

[17] It is related to, but by no means identical with, what others (e.g. Michael Frede) have called doxography.

From the Doxological point of view, the past of Philosophy is another vast continent where may be found philosophies not found closer to home which (because the work of developing them has already been done) can be imported for less than the cost of manufacture and, in principle at least, might compete with the local products. They can be assimilated. Such assimilation involves examining such philosophies with the sorts of considerations and the sorts of weighing of such considerations that one would use in constructing or examining some aspect of a contemporary philosophy. Doxology has an enormous and legitimate appeal. As Ian Hacking points out,[18] in a given milieu some texts speak to philosophers and their students with an immediacy with which others do not. That immediacy is largely a matter of their appeal as philosophy. These texts raise issues, ideas, and arguments that form the material for Doxology.

Doxology has its limitations. Precisely because it is not relevant to the doxologist as such whether a text is from Philosophy's past or present, there is a constant danger of anachronism—of missing what is in a text by failing to note its diferences from contemporary Philosophy. This threat of too close an assimilation to contemporary Philosophy is one major philosophical fuel of the Anthropology of Philosophy, which seeks to study past philosophies in their native habitats. The Anthropology of Philosophy is particularly important to philosophers dissatisfied with current normal practice. Such Philosophers seem to take one of several routes. Some, like Descartes, portray themselves as a new beginning, but others like Farabi or Heidegger, reach back to Philosophy's past, and present themselves as taking up an earlier and in some sense 'lost' tradition. Some of these (like Heidegger) do so by creating a new terminology and framework which would have been foreign to the past they attempt to recapture. Others attempt to immerse themselves and those they influence in the terminology and framework of the past they are attempting to recover—to see some part of that past as those for whom it was present saw it. This effort to see the past as the past saw itself when it was present is what I call Anthropology of Philosophy.[19] At its most banal it is a paraphrase of texts in their own vocabulary. Just as there are some texts from the past that 'resonate' with us because we see in them projects and considerations akin to our own, so there are other texts which speak with a different immediacy, seeming intriguing but mysterious, and suggestive of ways of thinking alluring but different from our own. These form the core material for Anthropology of Philosophy.

Much of the interest the past of Philosophy holds for philosophers can be satisifed by Doxology and Anthropology of Philosophy, and any study of the past of Philosophy which is going to interest philosophers qua philosophers will consist largely of these subjects. Given this, and given the fact that the past of Philosophy is studied largely among philosophers, it is not surprising that most of the work done on the past of Philosophy is in fact of these two kinds. So should it be!

[18] In I. Hacking 'Five Parables' in Rorty, Schneewind, and Skinner, eds. *Philosophy in History*, p. 106.

[19] Doxology and Anthropology of Philosophy correspond roughly to the two modes of studying the past of Philosophy canvassed by Richard Rorty at the beginning of 'The HIstoriography of Philosophy: Four Genres' (in Rorty, Schneewind, and Skinner, eds., *Philosophy in History*, pp. 49–76. However, while Rorty seemed to see them as competing modes I thinks they compete only when presented as complete approaches to the past of Philosophy. As I see it the History of Philosophy employs both.

For neither Doxologists nor Anthropologists of Philosophy is it crucial that what they study is of a different time (though often such study is embedded in mythoi about the past). One can (and perhaps some do) study Descartes' *Meditations on First Philosophy* as if it were written just a few years ago. Perhaps when the Jesuits encountered China (that 'nation of philosophers') some may have practiced what I call Anthropology of Philosophy in attempting to understand then current Chinese thought in its own terms. Why then might philosophers be interested in the past of Philosophy as *past*?

Both the Doxological and the Anthropological perspectives are partial. Because the doxologist does not study the past of Philosophy historically, to the extent that what is meant by the formulations offered by past philosophers can be best (or only) understood by seeing them in the context of the dialectical debates in which those philosophers were engaged, doxologists will fail to understand them and will assimilate them too closely to theses and arguments their contemporaries might offer. Moreover, if Philosophy is indeed a 'human kind', both the concepts employed in contemporary philosophy and the self-conception of contemporary philosophy itself have been and continue to be shaped by our conception of them. If, as seems likely, mere possession of concepts does not involve full mastery of them, we enrich our understanding of our own philosophical views by enriching our grasp of those concepts, and we can do this by coming to see more clearly how they have evolved.

Anthropologists of Philosophy, on the other hand, are in danger of forgetting their audience. If one does not already understand the language, just having it repeated louder and more slowly is not likely to help much! Should it happen that a past philosophical milieu was revived and became the milieu of contemporary philosophy, Anthropology of Philosophy would become Doxology. I don't know that that ever has happened—though certainly neo-Scholasticism was an effort in that direction!

These considerations suggest that, sometimes at least, to understand a text in a satisfying way one needs both Doxological and Anthropological approaches, and, since it is widely agreed that many of the best philosophers are dead, good reason for philosophers who wish to dialogue with the best philosophers on whatever terms to be interested in dialogue with the dead.[20] Since the living have difficulty getting responses to questions and arguments from the dead, there is a role for Historians of Philosophy to suggest how they would or might respond to contemporary concerns. To do it responsibly they must determine what the dead philosopher actually thought and what philosophical grounds the dead philosopher might have had for thinking it.

From the viewpoint of someone interested solely in developing a philosophical account of some particular phenomenon, the practice of Doxology and Anthropology of Philosophy might be enough, but from the viewpoint of a philosopher seeking to understand Philosophy itself, more may be needed. Doxology and Anthropology of Philosophy, even taken in tandem, are limited approaches to the past of Philosophy, even from a philosopher's perspective. Both take for granted a canon of philosophers and texts generated by particular Philosophical communities. A serious question a philosopher might ask is why that canon has the status it does. Philosophers are interested in the work of other philosophers, but no one can read, let alone think seriously about, the work of everyone else. Why have some figures and some texts become the foci of attention so that they are

[20] See E. Curley, 'Dialogues With the Dead' (*Synthese* 67, April 1986), pp. 33–49.

generally acknowledged as important and can, for example, form the basis of a more or less standard curriculum? Why, for example, did Porphyry decide to make the study of Aristotle's *Organon* a preparation for Platonic metaphysics and why did others follow him? Why was all of Aristotle and only a little of Plato translated into Latin in the twelfth and thirteenth centuries? Why are Descartes and Locke standard figures in courses in Early Modern Philosophy and not Hobbes (save in Political Philosophy courses) or Desgabets or Samuel Clarke? A philosopher might seek to answer such questions by alluding to what are, or were, considered philosophical considerations.

Again because Philosophy largely proceeds by conversation and debate (face to face or not) and the framework for such conversation and debate is very largely set by the issues and considerations already set out by others, questions about why it is those issues and considerations that are being discussed are serious philosophical questions. Some philosophers (Michel Foucault, Alastair MacIntyre, Richard Rorty, Quentin Skinner, and Charles Taylor come to mind) use narratives about the past of Philosophy to support the claim that fundamental posits of some large segment of contemporary Philosophy (for example, that thought is a matter of generating and manipulating representations) are dispensable. They do this by arguing that these posits, far from being forced upon us by the nature of things, arose in particular historical circumstances for contingent reasons and can well be done without. Since this is a historical claim, one might look to the History of Philosophy to determine its truth. Were it true, we would have philosophical grounds for philosophizing differently. One might well think then that knowing the history of the concepts assocated with contemporary issues would have philosophical consequences.

There is another, more meta-philosophical reason why philosophers might be interested in going beyond Doxology and Anthropology of Philosophy. Besides looking to the past of Philosophy for models for its future, philosophers often look to the past of Philosophy for their conception of Philosophy itself. One can philosophize today by taking up issues in Aristotle or Hegel or Locke, but not by taking up themes from Euclid or Hippocrates. This is because what counts as Philosophy is shaped by a communal perception of who were philosophers and what the boundaries of their philosophy were. One way of doing Philosophy is to challenge this perception and one way to challenge it is to retell the history of Philosophy in a new way, but any such retelling is against the background of a 'standard story'.

The considerations just adduced for a philosopher to be interested in a subject we might call the History of Philosophy as contrasted with either Doxology or Anthropology of Philosophy have a common structure: they are all considerations arising from reflection upon, rather than involvement in, current Philosophical practice. This suggests a thesis: Philosophers need take to heart the History of Philosophy (as contrasted with Doxology or Anthropology of Philosophy) when dissatisfied with current Philosophical practice. The special role of the History of Philosophy within Philosophy is to be a tool for use when Philosophy itself is in crisis.

Whether or not there is at any given moment a crisis within Philosophy is as controversial as the nature of Philosophy itself. So far I've been writing as if everyone who comes to the History of Philosophy comes to it excited about one or another of the current incarnations of Philosophy, but this is not so. A significant minority of scholars (some of whom can be heard at conferences announcing such theses as that Metaphysics died in 1781) come to it from a disappointed love of Philosophy. Having dreamt dreams not

satisfied by what they find in Philosophy's present, they turn to its past in hopes of finding its glory and explaining its decline. This turn to the past is a complex affair. On the one hand, it involves what I've been calling Anthropology of Philosophy—an excavation of concepts and modes of reasoning that have fallen out of favor and may now be only dimly understood. On the other, it involves locating these in a narrative of how Philosophy came to turn away from them. Since this turning away is characterized as a mistake, this involves a discussion of why the considerations which led to the abandonment of the path that is now being resurrected were inadequate ones. This in turn involves a narrative of change in Philosophy.

This brings us to the role of History in History of Philosophy and that of History of Philosophy in History. That an internal history of Philosophy is possible depends on it being the case that Philosophers adopt and change views on the basis of philosophical considerations and reasons. That is itself a controversial philosophical position. Many of those who would have Philosophy be very different accept it, and to the extent that they find in the history of Philosophy grounds for philosophizing differently, they will have to present those grounds as considerations and reasons which might appeal to us. Others outside Philosophy need not be so restricted, however.

Historians and sociologists are (sometimes) interested in what shapes and is shaped by the work of philosophers. Sometimes they might agree that what does this shaping is itself internal to the practices of Philosophy. Sometimes, however, they might very plausibly insist that what shapes and is shaped by the work of philosophers is well outside Philosophy. At one extreme we find this approach in the work of Randall Collins.

Randall Collins' remarkable *Sociology of Philosophies* is an effort to see the entire past of Philosophy very broadly. Collins' aim is simply stated: 'I am arguing that if one can understand the principles that determine intellectual networks, one has a causal explanation of ideas and their changes.' [21]

Collins' central claim is that it is intellectual networks and neither individuals or reasons as I have gestured to them above that are the propelling force of the history of Philosophy. The generation and sustaining of networks, he argues, is best understood not in terms of the content of the philosophies involved but in terms of *interaction rituals* in the sense of Irving Goffman. Philosophers interact with one another in ways that increase or decrease *emotional energy* and *cultural capital*. As he puts it:

> The social structure of the intellectual world, the topic of this book, is an ongoing struggle among chains of persons, charged up with emotional energy and cultural capital, to fill a small number of centers of attention. These focal points, which make up the cores of the intellectual world, are periodically rearranged; there is a limited amount of attention that can be distributed through the total intellectual network, but who and what is in those nodes fluctuates as old intellectual movements fade out and new ones begin. These nodes in the attention space are crescive, emergent; starting with small advantages among the first movers, they accelerate past thresholds, cumulatively monopolizing attention at the same time that attention is drained away from alternative nodes . . . My sociological task is . . . to see through intellectual history to the network of links and energies that shaped its emergence in time.[22]

[21] Randall Collins, *The Sociology of Philosophies* (Harvard University Press, Cambridge, MA., 1998), p. xvii.

[22] R. Collins, *The Sociology of Philosophies*, pp. 14–15.

Collins does not deny that Philosophers respond to reasons or that they employ concepts and arguments. His claim is rather that it is in virtue of their effect on the 'emotional energies' of philosophers that reasons and arguments are historical forces, and hence one can abstract away from the particular content of philosophies and focus instead on the ways these emotional energies are generated and dissipated. Philosophers gain emotional energy in having their proposals taken up by others and lose it in having them dismissed. Such abstraction, he suggests, shows us patterns of interaction which explain intellectual change. Historically important Philosophers (measured by how often they are cited by others) are members of intergenerational networks of high energy—they almost always have had Historically important Philosophers as teachers and mentors and often have other Historically important Philosophers as students or disciples. One striking example of the power of such abstraction is Collins' 'law of small numbers'—that the number of prominent Philosophical movements in a given milieu over any but the shortest periods of time will always vary between three and six. This has the consequence that if there are more than six there will be a tendency to consolidate, and if there are fewer than three a tendency to fission. This is independent of the content of the movements in question.

Collins' approach may illuminate the past of Philosophy. If he is right there is a pattern to the way philosophical movements arise, flourish, and decay that is repeated over and over again wherever anything recognizable as Philosophy is present. Indeed, Collins' arguments suggest practical advice to would-be philosophers who wish (as he suggests they normally do) to be influential—that they locate themselves within networks of influence—seek out influential teachers, take on talented students, and position themselves relative to contemporaries in the slots suggested by the Law of Small Numbers.

Nonetheless, there is an important sense in which Collins' results are irrelevant to the History of Philosophy. No text or argument or thesis is illuminated by them, and to the extent that changes in philosophical views are explained by social facts about Philosophical networks such changes (as contrasted with the fact that there *are* such changes) are of no philosophical interest.

Even if one's approach is less general than Collins', one might think that the past of Philosophy is illuminated by considerations from without it. One might, for example, study the impact of developments in the Christian conception of the Eucharist on Medieval Metaphysics or the impact of eighteenth-century Political Philosophy on the French Revolution. One might study the structural similarities among Philosophical communities. To a historian the discovery, development, acceptance, and rejection of philosophical views are historical events, and since for a historian to understand an event is (at least in part) to understand its causes, whatever causes are at work must be studied—whether or not they are plausibly regarded as reasons or as internal to the practice of Philosophy. Similarly, the categories and concepts a sociologist of knowledge might be expected to bring to bear on the study of Philosophy and its past are unlikely to be uniquely philosophical. What then might we think such a historian to miss that an Historian of Philosophy would not? It is, I suggest, precisely the role played in the development of Philosophy by the reasons and considerations the giving and taking of which constitute much of Philosophical practice.

As a philosopher one may be interested in, say, Aquinas' 'Third Way' and one might gain a deeper understanding of the argument both by comparing it with other texts in Aquinas himself and by studying and comparing it with arguments one finds

in Aristotle, but there is nothing historical about that other than that Aquinas and Aristotle are both dead. As a historian one attempts to explain why things were as they were at one time in terms of their relations to how things are at another. One might, for example, try to explain how Aquinas' formulation of his Third Way came to be as it is by appealing to the influence on his thought of his understanding of Averroes' understanding of Aristotle's argument in *Metaphysics* XII.6. Qua philosopher one would be interested in the influence of Aquinas' understanding of Averroes only if it shed light on the argument itself. Qua historian one is interested in the argument itself only insofar as its content sheds light on such matters as how Aquinas came to hold or present it; how Averroes influenced Latin Philosophy, or how Aristotle came to be (mis)understood. Qua Historian of Philosophy one is interested in both and (when things go well) attempts both to illuminate the argument itself and to illuminate the roles it played. At bottom is a presupposition that the argument played the roles it did in part because of its virtues (and defects) as an argument.

The History of Philosophy is thus poised, somewhat uneasily, between Doxology and Intellectual History. Its balance depends upon the truth of an assumption—that the history of Philosophy can be profitably understood as an internal history in the sense that what accounts for the historical development of Philosophy is (the consideration of) philosophical issues. Tightly bound with this is the claim that Philosophical views and arguments are accepted and rejected on the basis of reasons. Philosophers qua philosophers accept and reject arguments and views on the basis of reasons and so if the development of Philosophy is to be understood internally it will have to be understood as a matter of coming (or not) to appreciate certain reasons.

I suggested earlier in this chapter that the Historian of Philosophy must tell the history of Philosphy in terms of considerations which are or were regarded as philosophical. What exactly is involved in a reason or argument being one that the present or a past Philosophical community would think philosophical? It is not required that the reason or argument be one that emerged in a Philosophical community. For example, the view that a part of a collection may have the cardinality of the whole may have emerged in Mathematics (over a long tradition of Philosophical objection) though both the position and the arguments for it are now part of Philosophy. Again the power of the combination of random alteration and some principle of selection may have been explored first by biologists but is now understood (and largely appreciated) in Philosophy. What is required for a reason or argument to be philosophical is that it *not* be one that is accepted by the Philosophical community simply because the members of that community are also members of another or because of the prestige of another—it is that it be one the community thinks is Philosophical. What counts as a Philosophical reason may change both because whether it is Philosophical may change and because whether it is counted as a reason may change. There is nothing in the History of Philosophy which requires supposing that either is invariant across time or space.

This opens one avenue for context to play an important role in the history of Philosophy. It may be argued by Historians (whether of Philosophy or not) that in a given historical setting something counted as a reason or as a part of Philosophy which did not so count in other settings. Discovering that other thinkers took some item as a reason might well provide evidence that Philosophers did too and so reorient one's understanding of (for example) some part of an uncontroversially Philosophical text. Discovering that other thinkers

took a text (say) to be part of Philosophy might provide evidence that it was and so that Philosophy was not quite as one had thought it was.

Still, there are limits to be observed here. One other discipline which has often loomed large in the thinking of philosophers is Theology, and it has often happened that philosophers adopted for Theological reasons positions that one might well have adopted on the basis of then prevalent Philosophical considerations but that the philosophers in question did not adopt on the basis of such consderations.[23] The study of such background conditions is a significant part of Intellectual History but only a peripheral part of the History of Philosophy. Indeed, it is only part of the History of Philosophy at all in the sense that the Historian of Philosophy notes it and goes on. Of course it may not be easy to tell, even when one knows why a thinker accepted or rejected a position, whether the reasons are or were taken to be philosophical. For example, the doctrine of 'real accidents' dismissed by Descartes had proved very useful in articulating the standard Latin Medieval doctrine of the Eucharist, and Ockham adduces no other grounds for it. Buridan, on the other hand, exploits it in his accounts of quantity and motion. The doctrine itself has contemporary philosophical interest since it seems in many ways close to current theories of tropes. There is little doubt that discussion of it forms part of the History of Philosophy *and* of the History of Theology—though specialists in each may take it in different directions.

The History of Philosophy is thus an odd and rare beast. It is odd because of its service to two masters. It is rare because neither master usually needs the whole range of its services and so they are rarely provided. What will be taught in Philosophy departments and schools will usually be just the doxological or the anthropological fragment and what will be taught (if anything is) in History departments and schools will usually be just that narrative that the Historian of Philosophy would note before passing on to study the philosophical considerations underlying it.

Coming (or not) to appreciate reasons is a psychological matter and likely to be influenced by all sorts of cultural, psychological, and sociological factors. These are, nonetheless, irrelevant to the History of Philosophy which instead presupposes that the thinkers in question respond to reasons *rationally* so that philosophical views are put forward as, and are, *reasonable* solutions to philosophical problems, and that changes of view are *reasonable* responses to problems perceived in earlier views. Although Historians of Philosophy need not suppose that what counts as rationality or as reasonable has remained the same every when and where, they must be prepared to determine whether a response is rational or reasonable according to whatever standard is in play. I suggest this responsibility is twofold. Such evaluation must take place both from the Historian of Philosophy's own perspective (typically that of the Philosophical community to which s/he belongs) *and* from the perspective of the thinker whose thought is being studied and that thinker's Philosophical community. The extent to which they will agree is in part an empirical question—one must be open to the possibility that standards of reasonableness as well as style of reasoning have changed. One must be open as well, though, to the possibility that they have not, and that to the extent that the same issues are under consideration in

[23] A prominent (and excellent) philosopher once asserted to me that it was false that one ought to be a vegetarian since Christ ate meat at the Last Supper. Interesting though this argument is, it is not (on the face of it anyhow) a philosophical argument, and the fact that this philosopher maintained or advanced it is not part of the History of Philosophy.

different epochs, later philosophies avoid some of the problems besetting earlier ones. This need not entail that later philosophies are on the whole more adequate than earlier ones because it is compatible with later philosophies turning out to have solved problems in earlier ones at the price of incurring new difficulties not raised by their predecessors, but it does entail that there is some internal pressure within the History of Philosophy to suppose that Philosophy progresses and that where it does not, either new issues are being considered, what counts as reasons has changed, or the causes are irrelevant to (or are merely to be noted as background conditions by) the History of Philosophy as a discipline.[24] It follows then that as accounts of the past of Philosophy, Histories of Philosophy will almost certainly be incomplete.

This incompleteness is not a defect. Histories are, of necessity, incomplete because the historian must select among the infinity of facts those whose presentation will illuminate for the purpose she has. The Historian of Philosophy selects facts which illuminate the (broadly speaking) rational relations among propositions and so provide reasons for holding or rejecting theses, or thinking or acting in certain ways. If a philosopher holds or rejects a thesis or thinks or acts in a certain way for reasons that have nothing to do with such relations there is nothing the Historian of Philosophy can do save point to some other sort of History which might have better luck.

The assumption that philosophers propose and reject what they do for philosophical reasons underdetermines how these proposals and rejections are expressed. Just as it is a plausible but defeasible assumption generally that things seem as they do because they are that way, so it is a plausible but defeasible assumption of the History of Philosophy that a text means what it seems to mean and that a philosopher accepts what that philosopher asserts. Just as various sceptical traditions have challenged the general assumption, so the more particular one has been challenged within the History of Philosophy, notably (but by no means exclusively) by Leo Strauss and his followers.[25] The general form of such challenges begins from the observation that there are contexts in which a reasonable person who believed p would not assert p and might, while believing p, assert even not-p, taken together with the claim that we should apply a very different set of interpretative techniques to that philosopher's work than we would otherwise. Given the observation, the question arises whether we can reliably determine by consideration of the circumstances in which a philosopher is writing whether it is such a context. The sceptic would have it that we cannot and so should practice a 'hermeneutics of suspicion'. The less sceptical theorist influenced by (say) Strauss, will claim that we can, that a large portion of the past of Philosophy consists of such contexts, and hence that much of what philosophers have written cannot be

[24] The issues here are very similar to those raised in *The History of Science* by Thomas Kuhn and his successors. That Philosophy does not always progress is fairly obvious. The most glaring case is the situation in Western Europe between (say) Boethius and (say) Anselm of Canterbury. The loss of much of the philosophical writing of the Greco-Roman world and the disappearance of a fairly large elite that could read philosophical Greek combined to bring it about that much of the Western European Philosophies of (say) the ninth century CE. were, as far as we can tell, far inferior to those of late antiquity. The social and economic changes which brought it about that early medieval Western European writers did not have many texts their predecessors did, may well explain why (say) Anselm does not consider (and so does not respond to) Aristotle's critique of Plato's theory of Forms, and may well be mentioned in a History of Philosophy, but they will be mentioned only to indicate a gap.

[25] See Leo Strauss, *Persecution and the Art of Writing*, University of Chicago Press, Chicago, 1952.

read on its face. Given the claim then, one should read such texts with what Paul Ricoeur has felicitously called a 'hermeneutics of suspicion'.

This challenge in its most general form strikes at the heart of the History of Philosophy because if it is the case that philosophical texts are 'coded' in a way that one can understand only by knowing in detail the social and political circumstances in which they were written (or later read), then the autonomy of the development of Philosophy is threatened in a much more radical way than it is by the fact that philosophers often take up issues or considerations that arise outside Philosophy. That latter fact is compatible with the claim that these issues and considerations only become part of Philosophy once they have been integrated into the practices of the Philosophical community. The challenge, however, has it that the core practices of the community are hostage to wider social forces.

The first thing to be said is that from the point of view of the Historian of Philosophy it usually does not matter if a philosopher meant what he said, because usually it is what was said that has been taken up in subsequent Philosophy. There are exceptions of course. For example, if Strauss himself reads Farabi or Maimonides as not meaning what they said and his interpretation is taken up, then for those who interpret it that way, it is not the text but the 'coded' text which enters into the History of Philosophy. Note though that whether it is the text or the 'coded' text that is taken up depends on what is made of the text and not on the text itself. A misreading of a text can be as historically important as a correct reading! It remains, though, that unless the misreading is itself philosophically grounded it is not a part of the History of Philosophy but a limit of it. That said, there remain the tasks of understanding the text, understanding the thought of the philosopher who wrote it, and understanding how each was later understood. All three are tasks for the History of Philosophy.

A reason for doubting that there is History of Philosophy as I've described it is the scope of the pretensions of the subject. For most of its history, Philosophy has purported to be the most general of all disciplines, seeking an understanding of how, in Wilfrid Sellars' phrase, things in the most general sense, hang together in the most general sense. Reasons for thinking that things hang together might arise anywhere in the vast terrain of things people think about, and so one might plausibly think that there can be an internal history of Philosophy only if 'philosophy' is just a name for whatever anyone thinks about—that is, if History of Philosophy collapses into Intellectual History (or perhaps just History)!

It is, however, not so. Philosophers may indeed take up considerations from anywhere, but considerable work is required to make them part of Philosophy and it is only once they become part of Philosophy at one time that they become part of the History of Philosophy at another. The mathematical problem of providing a theory of irrational numbers seems once to have been part of Philosophy and the proposals of Greek mathematicians about it may well be part of the History of Philosophy. Current theories of irrational numbers seem to play no role in Philosophy and, on current form, will play no role in its history. On the other hand, concern about the ontological status of numbers seems to have played a significant role in ancient Philosophy and also does in contemporary Philosophy, while having played a very minor role, it still seems, in Medieval Philosophy. Historians of Medieval Philosophy can and do pay much less attention to Medieval Mathematics, not only because there seems to been less of it but because it seems never to have been taken up in a community then or now recognized as Philosophical. Again, while Shakespeare's plays contain much worth thinking about and may come to have a role in Philosophy (and thus in its

history) they do not now form part of that History and need not be studied by the current Historian of sixteenth-century Philosophy. Of course none of this proves definitively that there is or ever was a distinctive Philosophical community. We may discover that to understand Leviathan (or some other text we take or they took to be philosophy) we are well advised to think about Henry VI but that case would have to be made in detail. Meanwhile those of us attempting History of Philosophy have not reason to change our ways.

Because the History of Philosophy as I understand it is poised between Doxology and Intellectual History it can be approached from either. On the one hand, Historians of Philosophy might seek to understand the history of some part of Philosophy's past by asking the questions and applying the standards for answering them that Philosophers now employ. On the other hand, they might seek to understand that past in terms of the questions being asked and the standards for answering them being employed in the larger contexts of the period being studied. This Janus-faced nature of the History of Philosophy has the consequence that even among those who practice the History of Philosophy as I have described it there is little consensus on its method. Some eminent practitioners emphasize intellectual, cultural, and (even) social context in an effort to grasp what thinkers in the past would have taken to be good reasons and good arguments. Others emphasize the interaction with Philosophical issues in our own milieu. As I see it both are useful responses. They agree in focus on the issue of what are good reasons and good arguments. They disagree about where to look to settle that issue.

3. Practice

The History of Philosophy as I have described it is a demanding subject. The ideal Historian of the whole of Philosophy would be someone who knows (inter alia!) Greek, Latin, Arabic, Persian, Hebrew, English, French, German, Italian, Sanskrit, and Chinese, is a good philosopher who is fully conversant with contemporary Philosophy in its various incarnations, is a decent mathematician, has a wide knowledge of several literatures, and has at least very mild super powers. Such beings are rare. What are the rest of us to do? I propose some maxims:

3.1 Specialize

One thing we do (and should do) is specialize. We study (say) Aristotle or Descartes or Locke or Hegel or we study 'Ancient Philosophy' or 'Early Modern Philosophy'. Historians of Philosophy who specialize in Aristotle will (as matters stand) be forgiven an ignorance of Arabic, Persian Hebrew, Sanskrit, and Chinese (and in some quarters even Latin) because it is now supposed that nothing written in these languages will shed light on what Aristotle thought or what the texts we have from him mean. Such specialists will be expected to have a passing knowledge of the cultural, political, and social conditions of the world of Aristotle's time but not a special expertise in them, because it is assumed (indeed as Aristotle himself suggests!) that Aristotle was reacting not to them but to his Philosophical

predecessors. They will be expected to read and react to the recent commentary literature but large swaths of earlier commentary literature may be (and usually will be) ignored on the assumptions that it has in some sense been superseded. Most of all they will be expected to study the texts of Aristotle we have, to have large portions of them 'in mind', and to be able to speak to us about what they mean and how they relate to one another. In short, they will be expected to practice Doxology and, to some extent, Anthropology of Philosophy. Qua specialist in Aristotle they may not be expected to have any particular historical views—even say about his relation to Plato or Theophrastus—though quite typically they will.

3.2 Do Doxology

Doxology is and will remain central to the practice of the History of Philosophy as long as it is centred in Philosophy departments and schools. That is a good thing. At any given moment there are texts of the mighty dead which appeal immediately to students with an interest in philosophy (though which they are changes with time). In my own current experience the *Meno* and *Phaedo,* Boethius' *Consolation of Philosophy*, Ibn Tufail's *Hayy ibn Yaqzan*, Descartes' *Meditations*, Hume's Dialogues *Concerning Natural Religion*, Nietzsche's *Geneology of Morals*, and (usually) Russell's *Problems of Philosophy* can all be counted upon to hold the attention of undergraduates. Interpreting them (and many others) *to us* is not only pleasant, it is useful!

3.3 Aspire to Anthropology of Philosophy

Here is a test for one kind of understanding of a philosophy. Read to the bottom of a page you have not read or no longer remember. Stop and predict what you will find as you turn the page. If you can reliably do so there is a significant way in which you understand what is going on. This test is a test for a kind of understanding sometimes called *verstehen*—the kind one has of how to behave in one's own cultural or natural environment.

One can, of course, be surprised in one's own cutural or natural environment. One can fail to understand when to *tu* and when to *vous*, when 'no' means 'not now', and when it means 'not ever'. Such failures are akin to the Philosophical failing of overgeneralizing from a restricted diet of examples. The classic remedy is the historiographical principle 'Exhaust the Evidence!'

3.4 Exhaust the Evidence

The Historian's maxim 'Exhaust the Evidence' can be at best a regulative ideal. If we understand by evidence whatever bears on the settling of a question it is likely that there is no finite specification of *all* the evidence bearing on any question in the history of philosophy. Why were not more Platonic dialogues translated into Latin in the twelfth century? To such a question details of the life of Henricus Aristippus, details of the history of transmission of

Greek manuscripts, details of the many factors which shaped early medieval impressions of the relationship between Plato and Aristotle, as well as details of the Philosophical issues at stake in the Philosophical communities of the twelfth century and no doubt much else, will all be relevant. Even if, as I have suggested, of the factors named only the last is part of the History of Philosophy there may be no end of inquiry into those details.

'Exhaust the Evidence' is, however, at least a regulative ideal. It is an appropriate criticism of a Historian of Philosophy that s/he failed to consider a Philosophical factor that, when considered, changes the picture at issue. We are all, and always, open to such criticism (and generally eventually subject to it). The search for the grounds of this criticism—often in hitherto neglected texts—is an important source of progress in the discipline.

3.5 Struggle with Strauss

I suggested earlier in the chapter that it is a default presumption of the History of Philosophy that a good interpretation of a text—one that makes sense of it as a whole—is a good guide to the meaning of the author. What, though, when one cannot find such an interpretation, for example because the text on its face contradicts itself?

One approach to this situation, that is taken nearly universally by medieval commentators on Aristotle, for example, is to deny that texts ever are inconsistent and attribute the failure to find a satsifactory interpretation to the failings of the interpreter. A second, popular among late nineteenth- and twentieth-century Historians of both Ancient and Medieval Philosophy, is to claim that the author changed her/his mind and to deconstruct the text, arguing that it has distinct layers composed at different times. A third, less common approach is to argue that there are philosophical reasons (usually unexamined philosophical presuppositions) why the author did not notice the inconsistencies and a fourth is to claim that the inconsistencies are intentional. All of these approaches have difficulties, but the last, to suppose the author was deliberately inconsistent, is the one to be approached with the most caution—because to adopt it is to abandon the usual canons for determining the worth of an interpretation. Consistency may be the hobgoblin of small minds but it is the lifeblood of History of Philosophy.

4. Conclusion

I've stressed above that the History of Philosophy presupposes that the past of Philosophy is shaped by considerations of reason and that such considerations have a historical dimension. The situation here is perhaps usefully compared with the History of Science. The History of Science is also contested territory but it is territory that scientists have more or less abandoned to the Historians. When we learn that Newton spent as much time trying to decipher the Book of Numbers as he did trying to formulate a Mechanics, we learn, we are told, something important to the History of Science. No physicist, qua physicist, would or should care a fig, because important though that fact may be for understanding relations between physics and religion in the seventeenth century, it has no consequences (as

far as I know) for Physics. No physical theory or argument is affected by whether it is so or not. To the (limited) extent to which the History of Physics is studied in Departments of Physics the fact (if it be one) is an oddity. To the extent to which the History of Physics is studied in History Departments the fact may be taken very seriously inter alia as a sign that the History of Physics is just part of a larger cultural context, but it will almost always be taken seriously in conjunction with a relative neglect of what exactly Newton thought he was doing. [26]

The comparison between the History of Philosophy and the History of Science raises another issue. From the beginning of institutional History of Science there has been a strong movement maintaining that History of Science and Philosophy of Science are inseparable. At the core of the movement seems to be the conviction that understanding how modern science developed both shows us what it *is*, and enables the sweeping away of myths that have grown up about it. In particular, the work of T. S. Kuhn and those who have followed him has been taken to dispel the myth that what we think of (have thought of?) as scientific progress is a *rational* process. Kuhn's central claim that older scientific theories are not so much disproved as abandoned by young Turks with new ideas and reputations to be made, has suggested far-reaching philosophical claims—not least about the relativity of reason itself. Now Philosophy, of all subjects, is committed to self-reflection. If the History of Science is inseparable from the Philosophy of Science, one might think that the History of Philosophy would be inseparable from the Philosophy of Philosophy and, since the Philosophy of Philosophy is a branch of Philosophy, in that way inseparable from Philosophy itself. Thus it is that while the History of Philosophy may be in service to two masters, it is to Philosophy that it renders the greater service.

References

Bréhier, Emile, *Histoire de la Philosohie* tome 1 (Librairie Félix Alcan, Paris, 1928).

Collins, Randall, *The Scoiology of Philosophies* (Harvard University Press, Cambridge, MA., 1998).

Condren, Conal, *Hobbes, the Scribblerians and the History of Philosophy* (Pickering and Chatto, London, 2012).

Condren, Conal, Gaukroger, Stephen and Hunter Ian, (eds), *The Philosopher in Early Modern Europe: The Nature of a Contested Identity* (Cambridge University Press, Cambridge, 2006).

Crivelli, Paolo, *Aristotle on Truth* (Cambridge University Press, Cambridge, U.K., 2004).

Curley, Edwin, 'Dialogues With the Dead', *Synthese* 67, 1 [April 1986]), pp. 33–49.

Frede, Michael, *Essays on Ancient Philosophy* (University of Minnesota Press, Minneapols, 1987).

Hacking, Ian, 'The Looping Effects of Human Kinds', in *Causal Cognition,* ed. D. Sperber, D. Premack, and A. James Premack (Oxford University Press, New York, 1996), pp. 351–82.

Hadot, Pierre, *Philosophy as a Way of Life* (Blackwell, Oxford, 1995).

Hadot, Pierre, *What is Ancient Philosophy?* (Harvard University Press, Cambridge, MA, 2002).

[26] And, one suspects, as part of a campaign to cut the prestige of Physics down to size—but that is another matter!

Neiman, Susan, 'Meaning and Metaphysics', in *Teaching New Histories of Philosophy*, ed. Jerome B. Schneewind (Princeton University Press: Princeton, 2004), p. 40.
Normore, Calvin 'Doxology and the History of Philosophy', *Canadian Journal of Philosophy* Supplementary volume 16, 1990, pp. 203–26.
Piaia, Gregorio and Santinello, Giovanni, ed., *Models of the History of Philosophy* Volume II: *From the Cartesian Age to Brucker* (Springer, Dordrecht, 2011).
Rorty, Richard, Schneewind, Jerome B., and Skinner, Quentin, eds., *Philosophy in History* (Cambridge University Press, Cambridge, U.K., 1984).
Schneewind, Jerome B. ed. *Teaching New Histories of Philosophy* (Princeton University Press, Princeton, 2004).
Strauss, Leo *Persecution and the Art of Writing* (University of Chicago Press, Chicago, 1952).

CHAPTER 3

METHODOLOGY IN NINETEENTH- AND EARLY TWENTIETH-CENTURY ANALYTIC PHILOSOPHY

SCOTT SOAMES

1. Introduction

In the fall of 1910 and the winter of 1911, G. E. Moore gave a series of 20 lectures which were published 42 years later (in substantially their original form) as *Some Main Problems of Philosophy*.[1] The first lecture, "What is Philosophy?," is a useful indicator of the state of analytic philosophy in its early years. In it Moore discusses what he takes to be philosophy's most important questions, outlines competing answers, and points to what later lectures will make clear to be his own position on many of these questions. Looking back a century later, the contemporary reader can't help being struck by the thoroughly traditional conception of the aims of philosophy embraced by a founding father of a tradition that has often been seen as a revolutionary new departure in the subject.

For Moore, the most important task of philosophy is to give a general description of the whole universe—by which he means an accounting of the kinds of things we know to be in it (material objects, human minds, etc.), the kinds of things which, though not known to be in it, may very well be (e.g. a divine mind or minds, human minds after death), and the relations holding among the different kinds of things (e.g. minds *attached* to bodies). Related to this metaphysical quest is the epistemological task of explaining how we are justified in knowing most of the things we ordinarily take ourselves to know. Finally, Moore thinks, there are questions of value, the answers to which are independent of, and not derivable from, natural or metaphysical fact, including questions concerning the rightness or wrongness of actions, the goodness or badness of states of affairs, and even the value of

[1] Moore (1910–11).

the universe as a whole. In short, metaphysics, epistemology, and ethics (traditionally conceived) make up the core of his conception of philosophy.

Were we to supplement this sketch with the contemporaneous views of the two other major analytic figures of the day—Frege and Russell—metaphysical and epistemological questions concerning logic, language, and mathematics, viewed as inextricably connected, would be added to Moore's chief philosophical concerns. But the overall conception of philosophy wouldn't change much. In these early days of the analytic tradition some previously neglected philosophical topics were given new prominence, but they didn't replace traditional concerns, which continued to be addressed in new ways.

2. Frege's Logicism

The chief change was the rise of logical and linguistic analysis as the means to achieve essentially traditional ends. The great engine of innovation was logicism, which was motivated initially by two questions: "What are numbers?" and "What is the basis of mathematical knowledge?" It was Frege who led the way in answering these questions.[2] Convinced that the highest certainty belongs to elementary, self-evident principles of logic—without which thought itself might prove impossible—he believed that the sublime certainty of arithmetic and higher mathematics (save geometry), must be deductively based on logic itself.[3] It was to demonstrate this that he developed modern symbolic logic in his 1879 *Begriffsschrift*.[4] The key step after that was to derive arithmetic from logic by (i) specifying a small set of logical truths of the highest certainty to serve as axioms, (ii) defining all arithmetical concepts in terms of purely logical ones, and (iii) producing formal proofs of all arithmetical axioms from these definitions plus the axioms of logic.

The audacity of the program was partially mitigated by his muscular conception of logic. For Frege, logic carried its own ontology. An infinitely ascending hierarchy of predicates was matched by an infinitely ascending hierarchy of concepts they denoted. First-level concepts were functions from objects to truth values, second-level concepts were functions from first-level concepts to truth values, and so on. In addition, every concept had an extension (itself taken to be a kind of object), which we may regard as the class of entities (possibly empty) to which the concept assigned the value truth. Frege's "logical axioms" guaranteed the existence of multiple entities of this sort. Today many would say that his logic looks a lot like set theory, which is now widely regarded, not as logic per se, but as a fundamental mathematical theory in its own right.

But this is hindsight. The genius of Frege's philosophy of mathematics was his methodology for using his logico-set-theoretic foundation to address his deep philosophical questions about mathematics. Although prior to philosophical analysis we all know many arithmetical truths, we have no idea what numbers are and little understanding of how it is possible for us to achieve certain knowledge of them. *Frege's basic idea is that natural numbers are whatever they have to be in order to explain our knowledge of them*. Thus the

[2] See, Frege (1884).
[3] See section 14, and Preface, section xvii, of Frege (1893).
[4] Frege (1879).

way to discover what they are and how statements about them are justified is to frame definitions of each number, as well as the notion *natural number*, that allow us to logically deduce what we pretheoretically know from the definitions and other unproblematic knowledge.[5] How, for example, should 2, 3, 5, and addition be defined so that facts like those in (2) can be deduced from the definitions plus our knowledge of logic and empirical facts like (1)?

1. $\exists x \exists y$ (x is a black book on my desk & y is a black book on my desk & $x \neq y$ & $\forall z$ (z is a black book on my desk $\rightarrow z = x$ or $z = y$)) & $\exists u \exists v \exists w$ (u is a blue book on my desk & v is a blue book on my desk & w is a blue book on my desk & $u \neq v$ & $u \neq w$ & $v \neq w$ & $\forall z$ (z is a blue book on my desk $\rightarrow z = u$ or $z = v$ or $z = w$)) & $\forall x \forall y$ ((x is a black book & y is a blue book) $\rightarrow x \neq y$)

2a. The number of black books on my desk = 2 and the number of blue books on my desk = 3. (There are exactly 2 black books on my desk and exactly 3 blue books on my desk.)

b. The number of books on my desk = 5. (There are exactly 5 books on my desk.)

More generally, how might a proper understanding of what natural numbers and arithmetical operations are be used, first to derive our purely arithmetical knowledge from the laws of logic, and then to derive empirical applications of that knowledge by appealing to relevant empirical facts? This, for Frege, is the most important question that a philosophical theory of number must answer.

The fact that no other philosophy of mathematics of his day could answer this question was the primary basis for his devastating critiques of Millian naturalism, psychologism, and crude formalism (as well as for his dismissal of Kant's conception of arithmetic as founded upon our experience of time). His most fundamental objections to Mill and others were (i), that they often don't even attempt to answer this fundamental question, and (ii), that what they do say only gets in the way of a proper answer. Thus, he thought, they offered no real theory of number at all. By contrast his compelling conception of individual natural numbers as classes of equinumerous concepts (the extensions of which can be mapped 1-1 onto each other), his characterization of *the successor of n* (as the class of concepts G such that for some object x in the extension of G, the concept *being a G that is not identical to x* is a member of n), and his definition of the natural numbers as consisting of zero plus its ancestors under the successor function provided what appeared to be the basis for a powerful and elegant explanation of the knowledge to be explained.[6]

Not one to rest with appearances, Frege painstakingly performed the needed derivations in Frege (1903) which, unfortunately, could not stand the shock of Russell's paradox.[7] Though he cobbled together a temporary fix—which wasn't proven to be inadequate until many years later—he knew that the game, as he had conceived it, was up. The inconsistency

[5] See sections 46, 56, 60, and 64–6 of Frege (1884).

[6] The first two occurrences of 'G' in the above sentence are to be understood as occurrences of a variable ranging over concepts; the final occurrence is to be taken as a corresponding schematic letter. A bit awkward, but the idea should be clear enough.

[7] For a discussion of Frege's derivations see chapter 1, section 6, of Soames (2014); for a discussion of Russell's paradox and its threat to logicism see chapter 2, section 9 of that volume.

of his logico-set-theoretic system, plus the daunting task of repairing it without loss of requisite power carried an obvious lesson. Logico-set-theoretic axioms that had seemed to provide a bedrock of certainty sufficient to ground arithmetic and all of mathematics had been proven to be faulty. Although they could be, and eventually were, satisfactorily revised in various ways, the price of such revision was a loss of either expressive power or perceived self-evidence, or both. Either way, Frege's ambitious conception of the goals of the logicist project could not be satisfied. In the end, the task of *justifying* arithmetical and other mathematical knowledge in terms of self-evident logical principles for which no similar problem of justification could arise had to be given up. Simply put, the arithmetical axioms to be proved were, if anything, more secure and less in need of justification than the "logical axioms" posited to prove them. Hence, the classical version of the logicist project foundered.

3. Russell's Logicism

When it came time for Russell's version, both the logic and the philosophy had changed. Whereas Frege's higher-order logic involved ascending levels of quantification over concepts of higher and higher levels, Russell's is most naturally read as involving quantification over ascending orders of classes—individuals, classes of individuals, classes of classes of individuals, and so on. Whereas extensions of concepts—classes—are all available at the lowest level of Frege's system, in Russell's they come presorted, with all classes of elements of level n coming at the next level up. By dubiously treating the type restrictions regulating class availability as if they were constraints on the ability to speak meaningfully about classes at all, Russell was able to render his paradox unstateable, and so to preserve his system from contradiction. However, since the natural numbers had to be located at some specific level—they are classes the elements of which are classes of individuals—he needed his infamous Axiom of Infinity, which posits infinitely many individuals (nonclasses), to guarantee that he won't run out of numbers.[8]

Hence, the philosophy underlying his version of logicism had to change. Logicism, as classically understood, is the view that arithmetic and much of higher mathematics is derivable from pure logic, and so is properly a branch of logic itself. However, in Russell and Whitehead (1910) and later work, Russell recognizes that the Axiom of Infinity is at best an empirical, rather than a logical, truth—as well as being one he does *not* know to be true. In light of this, Boolos (1994) makes a good case that Russell's considered view was a weaker form of logicism according to which *mathematical concepts* are reducible to *purely logical concepts*, even though the proofs of many mathematical truths require non-logical existence claims about how many individuals there are.

This retreat, though significant, is not so bad—in part because it is extremely doubtful that the original, classical version of logicism is achievable, and in part because the weaker version contributes to goals different from the original aim of *justifying* mathematics. By 1907 Russell had come to appreciate the central difficulty with doing that. *We are more*

[8] See chapter 10 of Soames (2014).

certain of the axioms of arithmetic, and less in the dark about how we can know them to be true, than we are of the axioms of any purported system of logic or set theory to which they might be reduced. As he realized, his own theory of types plus axioms of comprehension and infinity raise more questions, and are subject to greater rational doubt, than the arithmetical system he derives from it. Hence the former can't be used to justify the latter.

By this time, he had come to view justification as going in the other direction—from the reduced theory to the reducing theory, rather than the other way. In Russell (1907) he argued that sometimes previously unknown and unobvious logical or mathematic truths can be justified by the fact that they provide *explanations* of the known and obvious truths that follow from them. The suggestion is that his unobvious logico-set-theoretic system is justified, at least in part, by the fact that the intrinsically obvious and antecedently justified theory of arithmetic follows from it.

As a principle of metaphysical methodology to be employed in developing philosophical foundations for mathematics, the principle that the metaphysically more fundamental may be justified by the explanation it provides of the epistemically fundamental is not unattractive. Some logical and metamathematical claims may be more foundational and explanatory than others, even though the latter may be more epistemically obvious than the former. When looking for the fundamental structure of a subject, one should, according to the view suggested, use the less foundational to justify the more foundational. But what exactly does this come to in Russell's case? In Russell (1907), he lays down three ways in which the derivation of arithmetic from his underlying system helps justify the latter. First, he says, in showing that the arithmetical axioms, and through them the theorems of classical mathematics, are derivable from his system, we see how our overall system of mathematical knowledge is (or can be?) organized, and how different parts of that system are related to one another. Second, he notes, the reduction can lead to useful extensions and unifications of mathematical knowledge, such as the extension of our ordinary notion of number to include transfinite numbers. Third, he claims that by illuminating the logical nature of mathematics we can throw light on the philosophical question of what mathematical knowledge amounts to, and how it is achieved.[9]

Although there may be merit in Russell's first two points, the third is more doubtful. Since his reduction relies on the Axiom of Infinity, which we don't, antecedently, know to be true, no appeal to it can explain how we achieved our antecedent knowledge of arithmetic and mathematics in general (supposing we have known these all along). Nor do we learn the Axiom of Infinity to be true by noting its role in Russell's reduction; he himself certainly didn't. So it is hard to see how his particular reduction succeeds in explaining anything about our knowledge of arithmetic. These issues could, of course, be sidestepped by carrying out the reduction in another way—for example, by deriving arithmetic from ZF set theory. But what would this accomplish? Is there some epistemic problem about arithmetic, and our knowledge of it, that is not equally a problem with ZF, and our knowledge of it? If so, I am not sure what it is.

That, of course, isn't how Russell saw things—in part because ZF was still on the horizon in 1910 and in part because he thought he had eliminated classes (sets) from his ontology. Although he allows himself to use the language of "sets/classes" he explicitly disavows

[9] Russell (1907), pp. 282–3.

commitment to them as entities. His general position is sketched in section 2 of Chapter 3 of the Introduction to Russell and Whitehead (1910).

> The symbols for classes, like those for descriptions, are, in our system, incomplete symbols: their uses are defined, but they themselves are not assumed to mean anything at all. That is to say, the uses of such symbols are so defined that when the *definiens* is substituted for the *definiendum*, there no longer remains any symbol supposed to represent a class. *Thus classes, so far as we introduce them, are merely symbolic or linguistic conveniences, not genuine objects as their members are if they are individuals.* (my emphasis, pp. 71–2)

The contextual definition of the usual class notation is given at *20.01.[10]

$$F\left(\{x:\ Gx\}\right) =_{df} \exists H\,[\forall y\ (Hy \leftrightarrow Gy)\ \&\ F(H)]$$

According to this definition, a formula that seems to say that F is true of the class of individuals satisfying G is really an abbreviation for a more complex formula that says that F is true of something that is true of all and only the individuals that satisfy G. According to Russell, this something is "a propositional function." If propositional functions were still functions from individuals to old-fashioned Russellian propositions, then to say that such a function is true of an object would be to say that it assigns the object a true proposition, and to say that propositional functions are *extensionally equivalent* would be to say they are true of the same things. (Similarly for two properties or for a property and a propositional function.) So whenever G and G* stand for extensionally equivalent properties or propositional functions, ⌜F ({x: Gx}) and ⌜F ({x: G*x}) will, according to the definition, agree in truth value—as they should.

Next consider a propositional function that takes a property or propositional function as argument. Call it *extensional* iff whenever it is true of its argument A, it is true of all arguments extensionally equivalent to A. As Russell notes at *Principia* *20, not all propositional functions are extensional in this sense.

> [the propositional function designated by] "I believe ∀x Φx" is an *intensional* function [and so not extensional] because even if ∀x (Φx ↔ Ψx), it by no means follows that I believe ∀x Ψx provided that I believe ∀x Φx. (p. 187)

Suppose that $p_{\Phi x}$ and $p_{\Psi x}$ are different but extensionally equivalent propositional functions, the former mapping an arbitrary individual a onto the proposition *that if a is a human, then a is a human* and the latter mapping a onto the proposition *that if a is a featherless biped, then a is a human*. Now let Φ be a first-level predicate variable. Then the propositional function designated by ⌜I believe ∀x Φx —which maps propositional functions onto propositions expressed by the corresponding belief ascriptions—may assign $p_{\Phi x}$ a true proposition about what I believe while assigning $p_{\Psi x}$ a false proposition. Thus, the propositional function designated by the belief ascription is *intensional*, rather than extensional. However, since, as Russell plausibly holds, *only* extensional propositional functions are relevant to mathematics, the system in Russell and Whitehead (1910) can be restricted to them. When one does this, the only thing about the proposition assigned by a propositional function to

[10] Russell and Whitehead (1910), p. 190.

a given argument that matters to the construction is its truth value. This being so, we can reinterpret the entire construction in terms of functions from arguments to truth values (rather than propositions)—*without losing anything essential to the reduction*.[11]

Although the result is pleasing, Russell would not have liked it. A function from arguments to truth values is *the characteristic function of the class* of things to which it assigns truth. There is no mathematically significant difference between working with classes and working with their characteristic functions; anything done with one can be done with the other. Nor does there seem to be any important ontological or philosophical difference between the two. But then, Russell's treatment of classes as "logical fictions" would have been empty and pointless—which is how most of the mathematicians, logicians, and philosophers who followed Russell saw it.

By contrast, Russell took the elimination of classes seriously. By the time Russell and Whitehead (1910) appeared, his view of propositional functions had become a radically deflationary version of his earlier "realist" view of them as nonlinguistic entities. The result was a thoroughly de-ontologized interpretation of his technical reduction. In this work he speaks of propositions and propositional functions in various, not always consistent, ways. But most of the time he seems to take propositions to be sentences, and propositional functions to be formulas one gets from them by replacing an occurrence of an expression with a free occurrence of a variable. Thus, looking back in Russell (1940) he says, "In the language of the second-order variables denote symbols, not what is symbolized" (p. 192), while in Russell (1959) he says, "Whitehead and I thought of a propositional function as an expression" (p. 92). If this is so, it would seem that a sentence of the form "∀P … P … " must mean that *every value of the formula "… P … "* is true.

Russell and Whitehead (1910) is replete with language like this. For example, in section 3 of chapter 3 of the Introduction, Russell sketches the idea of a hierarchy of notions of truth that apply to the different levels of his type construction. Assuming that truth has already been defined for quantifier-free sentences at the lowest level, he explains first-order quantification as follows:

> Consider now the proposition ⌈∀x Φx⌉. If this has truth of the sort appropriate to it, that will mean that every value of Φx has "first truth" [the lowest level of truth]. Thus if we call the sort of truth that is appropriate to ⌈∀x Φx ⌉ "second truth," *we may define ⌈∀x Φx ⌉* as *meaning ⌈every value for "Φx" has first truth⌉* . . . Similarly . . . *we may define ⌈∃x Φx ⌉ as meaning ⌈some value for "Φx" has first truth⌉*.[12]

Here, in addition to assuming that a similar explanation can be given for higher-order quantification, we are to assume that "first-truth" conditions and meanings have been given for quantifier-free sentences at the lowest level.

Although it might appear from this that Russell took quantificational statements to express metalinguistic facts about language (even though their instances make entirely non-metalinguistic claims), this surely cannot be right. There is, however, another interpretation that could be given to his remarks.[13] On this interpretation, the quantifiers in his

[11] See pp. 39–40, Burgess (2005).

[12] My emphasis, p. 42. I have here changed Russell's notation in inessential ways, and used corner quotes to clear up some of the use/mention sloppiness.

[13] This interpretation is defended in Landini (1998) and in Klement (2004).

reduction are what are now called "substitutional." So understood, they don't range over objects of any kind—linguistic or nonlinguistic, but rather are associated with substitution classes of expressions. Although their *truth conditions* are stated metalinguistically, *their content* is supposed to be nonlinguistic. Using objectual quantifiers over expressions, we can give substitutional *truth conditions* of quantified sentences in the normal way—as Russell does. ⌜$\forall x\, \Phi x$⌝ and ⌜$\exists x\, \Phi x$⌝ are true, respectively, iff all, or some, of their substitution instances are true, where the latter are gotten from replacing free occurrences of "x" in Φx by an expression in the relevant substitution class. This explanation will work, provided that the truth values of the sentences on which the quantified sentences depend are already determined before reaching the quantified sentences, and so do not themselves depend on the truth or falsity of any higher-level, substitutionally quantified sentences.

There are three important points to note. First, if one combines the hierarchical restriction inherent in substitutional quantification with Russell's system of higher levels of quantification, the type restrictions he needs for his logicist reduction will fall out from the restrictions on substitutional quantification, without any need for further justification. Second, on the substitutional interpretation, there is no need for what look like higher-level "existential" generalizations—i.e. ⌜$\exists P\, \Phi(P)$⌝, ⌜$\exists P_2\, \Phi(P_2)$⌝, etc.—to carry any ontological commitment. They won't—as long as the relevant substitution instances can be true even when the constant replacing the bound variable doesn't designate anything. Third, for this reason, it is tempting to think that no quantificational statements in the hierarchy carry any ontological commitments not already carried by quantifier-free sentences at the lowest level. Since Russell took accepting the latter to commit one only to individuals and simple properties and relations, it would be natural for him to characterize classes, numbers, and nonlinguistic propositions and propositional functions as "logical fictions," while nevertheless appealing to them when "speaking with the vulgar," as he does in Russell (1919).

The virtue of the substitutional interpretation of quantification is that it makes some sense of Russell's ill-advised enthusiasm for his no-class theory. It's vice is that it is technically insufficient to support the reduction in *Principia*, while undermining his other important achievements in philosophical logic. Although there are many points to be made along these lines, some rather complex, I will here briefly mention just one main point.[14] Russell's type theory involves quantification at each of its infinitely many levels. The first level includes what is now called "first-order quantification" over individuals. Though this can be simulated by first-order substitutional quantification, the price to be paid (required by Russell's Axiom of Infinity) includes positing infinitely many logically proper names of individuals. Though this is bad enough, a more far-reaching difficulty occurs with higher-order substitutional quantification.

At the next level we get second-order quantification. The associated predicates include all simple predicates used to construct atomic sentences, plus complex predicates. For any first-level sentence in which simple predicates occur, we need a complex predicate for each of the ways of abstracting one or more of the predicates via lambda abstraction—as illustrated by expressions like $\lambda F \lambda G[\Phi\, (\ldots F \ldots G)]$. All these predicates, simple and compound, are associated with the predicate variables. So, on the substitutional interpretation, ⌜$\forall X_1\, \Phi(X_1)$⌝ is true iff every sentence is true that results from erasing the quantifier and

[14] For a much more thorough discussion, see section 5 of chapter 10 of Soames (2014).

substituting an occurrence of a predicate, simple or complex, associated with "X_1" for each occurrence of "X_1" in $\Phi(X_1)$; similarly for second-order existential quantification.

Looking at this from the outside (where we continue to allow ourselves to speak of classes), this means that our substitutional construal of second-order quantification parallels ordinary objectual second-order quantification over *those classes that are extensions of first-level predicates of individuals* (including complex predicates). This process is repeated for third-order quantification, except that here complex predicates are the only ones in the substitution class. This level mimics objectual quantification over *those classes that are extensions of second-level predicates, members of which are classes of individuals that are extensions of first-level predicates*. (This is the level at which Russellian natural numbers appear as classes of equinumerous classes.) The hierarchy continues uniformly from there on.

It is important to notice the diminished expressive power of the substitutional versions of higher-order quantifiers. Whereas the objectual quantifiers range over *all classes at a given level*—both those that are extensions of predicates at that level (of the Russellian logical language) and those that are not—the substitutional quantifier mimics only quantification over the former. If, as is standardly assumed, every sentence and every predicate is a *finite* sequence of the logical and nonlogical vocabulary, the domain of all classes at a given level will far outstrip the domain of all classes that are the extensions of predicates at that level.[15] As a result, the expressive power of the underlying "logical" theory to which arithmetic is to be reduced is sharply diminished by treating its quantifiers substitutionally. Worse, this diminishment affects Russell's definition of *natural number*, and his use of that definition to prove the crucial arithmetical axiom of mathematical induction (which, informally, says that if zero has a property, and if whenever a natural number has a property, its successor does too, then every natural number has the property). As a result, we lose the simple Fregean way of understanding and proving the induction axiom.[16]

Now that we recognize the power of *objectual* second-order quantification, and its relevance for arithmetic (e.g. second-order arithmetic is complete whereas first-order arithmetic is not), the idea of depriving the higher-order system of *Principia Mathematica* of that power should be a non-starter. We certainly shouldn't do so in order to avoid ontological commitment to classes. Though Russell wasn't in a position to know this in 1910, classes have proven so useful for all sorts of theories—not least of which model theories for logic and semantics—that giving them up seems virtually out of the question. It is not even clear that we would know how to investigate the differences between first- and second-order logic, and first- and first- and second-order arithmetic, without classes.[17] To read *Principia*

[15] This can be avoided by allowing individual formulas to be infinitely long, and interpreting higher-order substitutional quantification in terms of substitution instances that are infinitely long. This strategy is meticulously developed in Hodes (2012), which shows it to be needed by a substitutional interpretation of higher-order quantifiers in *Principia Mathematic* leading to a ramified theory of types, without classes or non-linguistic propositions and propositional functions. As Hodes notes, the price of this approach is high, since it renders Russell's logical language incapable of being understood by agents whose cognitive powers are finite. See Gödel (1944) for related discussion.

[16] See section 5 of chapter 10 of Soames (2014) for details. For fruitful related discussion see Gödel (1944), pp. 145–6.

[17] Although there are new interpretations of second-order systems that use the notion of plural quantification, and although these interpretations have important uses, it is doubtful that they should be seen as replacing or eliminating set theoretic interpretations of second-order systems.

Mathematica in a way that distances him from the progress made in the tradition he helped to create would be to let tenuous philosophical thoughts about the elimination of classes—which were at best underdeveloped—obscure his positive contribution. That said, it must be admitted that some thoughts of a substitutional sort seem to have played a (confused) role in stoking his enthusiasm for his conception of logico-linguistic analysis as a method of ontological reduction in the philosophy of mathematics and beyond.

4. Russell's Logical Atomism

Russell's next significant steps were taken in Russell (1914, 1918–1919), in which two broad tendencies are discernable. The first is an ambitious analytic reductionism, by which he sought to avoid ontological commitment to entities thought to be problematic. Just as he took his theory of descriptions and his analysis of ordinary names as disguised descriptions to provide the treatment of negative existentials needed to finally put to rest his earlier broadly Meinongian ontology,[18] so he took his multiple-relation analysis of judgment to eliminate propositions,[19] and his reduction of natural numbers to classes to dispense with an independent category of abstract objects. But that was only the beginning. His more radical view of classes themselves as "logical fictions" advanced a strikingly minimalist metaphysical agenda, present from Russell and Whitehead (1910) onward. In Russell (1914), he took a further ontological step in the service of epistemological concerns. It was there that he renounced commitment to physical objects as independently existing substances, characterizing them instead as "logical constructions" out of the objects of immediate sense perception. By this time his view of reality had been stripped of all abstract objects except "universals"—simple properties and relations—and all particulars except for individual selves (themselves to be eliminated in Russell 1918–1919) and the fleeting, private objects of their immediate perception.

In all of this, we see the second broadly methodological tendency, which linked his evolving metaphysical minimalism with an ambitious search for secure epistemological foundations. At this point, Russell's epistemological practice consisted of two distinguishable sub tasks. The first was to isolate a domain of putative pretheoretic knowledge, which though revisable at the margins, was taken to be, on the whole, beyond serious doubt. In the case of logicism this domain was our knowledge of arithmetic and other branches of mathematics; in the case of the external world it was our knowledge of physical science, and of the truth of most ordinary judgments about "physical objects." The second sub task was to identify a minimal set of underlying notions to be used to identify analytically primitive judgments or axioms, plus definitions from which the overwhelming majority of the pre-theoretic claims taken as data could be analyzed/derived. Russell did not require the underlying axioms or definitions to be self-evidently obvious. It was enough that they could be used to explain how the pre-theoretic claims under analysis could be known to be true, while avoiding puzzles and paradoxes generated by the gratuitous postulation of entities the nature and existence of which we have no way of knowing.

[18] In, Russell (1905) and chapter 16 of Russell (1919).

[19] In Russell (1912).

Russell makes his methodology explicit in lecture one of Russell (1918–1919). His aim is to outline the structure of a world we are capable of knowing, based on the ideas (i) that most of what we are given by the empirical and deductive sciences, and by the most fundamental judgments of common sense, is true and capable of being known, (ii) that although we are justifiably confident of this, we do not know the real content of these truths, and (iii) that the job of analysis is to elucidate this content, thereby explaining how we do, or at least could come to, know it. He warns that we can't anticipate the end result of analysis in advance, and admonishes us not to dismiss what may seem to be highly revisionary characterizations of the knowable content of pretheoretic claims. At this point one wonders, is the task to explain what justifies the vast pretheoretic knowledge we already have, or is it to articulate what our evidence really justifies, and hence *what we could come to know, if we adopted a frankly revisionary view of the world*? Although Russell often speaks as if he adhered to the first conception, the results he reaches suggest the second.

Applying his method to physics, he says:

> You find, if you read the words of physicists, that they reduce matter down to . . . very tiny bits of matter that are still just like matter in the fact that they persist through time, and that they travel through space . . . Things of that sort, I say, are not the ultimate constituents of matter in any metaphysical sense . . . *Those things are all of them . . . logical fictions . . . It is possible that there may be all these things that the physicist talks about in actual reality, but it is impossible that we should have any reason whatsoever for supposing that there are.* (My emphasis, pp. 143–4)

Although this sounds startling—*There is no reason whatever to think that anything really persists through time or moves through space?*—it is simply a recapitulation of the view for which Russell argues in *Our Knowledge of the External World*. There he maintains that the only *knowledge* we use material-object sentences (from physics or everyday life) to express is *knowledge of our own private sense data plus the private sense data of others*. The task he envisions (but only vaguely sketches) is that of translating all material-object sentences into other statements (logical forms of the originals) in which only agents and their sense data are explicitly mentioned. It is only in this way, he thinks, that one can capture what science really teaches, along with what we *justifiably* report when we talk about things like tables, chairs, atoms, molecules, desks, or human bodies. Although he takes it to be possible to imagine epistemically inaccessible entities beyond our sense data, he takes it to be *impossible* to provide any empirical justification for such speculative claims, which are unknowable.

The new element in Russell (1918–1919) is his treatment of all sentient agents as *logical fictions*. Previously, he had regarded other minds as theoretical posits, the justification of which was that they were needed in the logical construction of material objects. Since, in the presence of his implicit epistemology, this move had all the virtues of an appeal to the unknowable Axiom of Infinity to "explain" our knowledge of arithmetic, he needed to eliminate commitment to minds from his mature atomism. On the view outlined in lecture 8 of Russell (1918–1919), each agent is a series of experiences that bear a similarity relation dubbed "*being experiences of the same person*." Of course series, along with classes, are *logical fictions* for Russell, so, no persons or other agents—in the sense in which persons or other agents can *justifiably* be believed to exist—really do exist. On the contrary, the only knowable particulars are momentary, private, sense

data. Some of these—call them "M-experiences"—are "mine," and others—call them "J-experiences" are "Jones's." All M-experiences bear the relation *being experiences-of-one-person* to one another, and to nothing else, while all J-experiences bear that relation to one another, and to nothing else (including the M-experiences). So, when I say that I exist, *knowing that what I say is true*, all I am really saying is that certain experiences are M-related. The same goes for Jones: when he says that *he exists* all he is saying is that other experiences are J-related.

The end result is a conception of (knowable) reality in which all particulars are momentary sense data bearing certain relations to other sense data. Allowing ourselves some useful "logical fictions" we can describe these particulars as arranged into two different cross-cutting systems of "classes"—those that constitute "agents" and those that constitute the things—like tables, chairs, human bodies, etc.—which we pretheoretically (but ultimately misleadingly) describe "agents" as "perceiving." To be sure, Russell didn't claim to know this fantastic conception of the universe to be correct. Still, it was his last pre-Tractarian stab at the truth. Who would have thought that supposedly hard-headed logical and linguistic analysis would lead to a metaphysical system as thoroughly revisionary of our ordinary conception of ourselves and the world presented to us in science or in ordinary life? More revisionary than Berkeley's system, Russell's analytic revisionism may surpass the dreamlike conception of Reality in McTaggart (1921, 1927) as an eternal, unchanging community of human souls. If one had thought that the new tradition in analytic philosophy had left such extravagant metaphysical speculation behind, Russell's example should convince one otherwise.

5. Epistemically Driven Metaphysical Minimalism

How did "logical analysis" lead to such a stupefying metaphysical vision? It did so because for Russell it was merely a tool in the service of a highly restrictive antecedent conception of what is required for empirical knowledge.[20] For the most part, he simply took it for granted without extended examination that perceptual evidence can't justify claims about genuine three-dimensional objects persisting through time which continue to exist whether or not they are perceived. In addition to this implicit presupposition about what knowledge must be, he was committed to an unusual sort of fidelity to most of our ordinary knowledge claims about ourselves, material objects, and the deliverances of science. Taking the pretheoretic content of these to be utterly opaque to us—much as Frege, more understandably, took our pretheoretic conception of number to be essentially blank—Russell required that *the sentences* used to express what we pretheoretically take to be empirically known to be capable of being systematically assigned contents that can be seen to be both true

[20] What conception is that? As a rule of thumb, it is the following: What the logical positivists were later to declare to be meaningless, because unverifiable, Russell at this stage merely labeled as in principle unknowable, and unjustifiable. This conception is discussed in chapter 11 of Soames (2014).

and justified by the restrictive body of sense experience he took to be available to us as evidence. In brief, a narrow and largely unreflective conception of what knowledge and justification consisted in, coupled with a commitment to speak with the vulgar, set the parameters for what counted as correct logic-linguistic analyses. "Analysis" yielded the results that it did, because, at this stage, it was merely a tool in the service of Russell's independent philosophical ends.

In assessing Russell's methodology concerning our knowledge of the empirical world, it is instructive to compare his approach to that of Moore's, as expressed in Moore (1909), Lectures 5 and 6 Moore (1910–1911), Moore (1925), and Moore (1939). When Moore took it to be evident that he knew that there were human hands, pencils, and what not, on the basis of perception, he did *not* mean merely that the *sentences* he used to report this knowledge could be assigned *some content or other* that would make them come out true and knowable. On the contrary, in Moore (1910–1911) and Moore (1925) he rules out phemonenalistic "analyses" of the sort advocated by Russell as *not* capturing the content of what he pretheoretically knows. The great failure of Moore's epistemology was, of course, that he was never able to make clear just how it is that perceptual evidence justifies what he rightly took himself to know.[21] However, this does not negate the crucial Moorean lesson: *Broad philosophical theories about knowledge are answerable to our firmest pretheoretic convictions involving particular things that we know, which cannot be reconstrued, or overturned wholesale, when they conflict with philosophical theory.*

It is also instructive to contrast Frege and Russell's transparent conceptions of meaning with their practice of assigning contents to sentences subject to philosophical analysis. When Frege presented his strategy for the logicist reduction, he contended that "the content" of sentences about number were, in effect, given by his ingenious set-theoretical translations of them. However, it was not clear in 1884 what his notion of "content" was supposed to come to, since this was well before his distinction between sense and reference in Frege (1892). Moreover, it is clear that Fregean translations of ordinary sentences of arithmetic do not share the Fregean senses expressed by the arithmetical originals—since an agent who understands both may sincerely assent to one without assenting to the other. A similar point holds for Russell's version of logicism. Given the usual criteria employed by both philosophers for determining when two expressions, or sentences, mean the same thing (think of Russell on logically proper names), the pretheoretically given sentences undergoing logicist analysis do *not* mean the same thing as the sentences that provide the analyses. This is all the more evident for his phenomenalist reduction of statements about the external world to statements about sense data. Thus there appears to be a sharp conflict between introspectivist accounts that take meaning to be highly transparent, seemingly favored by Frege and Russell, with some of their more ambitions philosophical claims about the fruits of logico-linguistic analysis. One way of dealing with this tension would be to take the results of "analysis" to replace, rather than explicate, the claims being analyzed. It is significant, however, that neither philosopher adopted this strategy forthrightly and consistently.

[21] For attempts to improve on Moore, see James Pryor (2000, 2004).

6. Logic and the Modalities

Up to now, I haven't said anything about modality. Since the distinction between necessary and contingent truth is mostly irrelevant when it comes to mathematics, it is not surprising that Frege and Russell were not concerned with it when developing their logicist views. However, when the focus of loco-linguistic "analysis" shifts to the external world, the distinction between what could and could not be, as well as between what would, versus what wouldn't, be if various conditions were fulfilled, becomes relevant. Although Russell paid very little attention to these issues, he did not ignore them entirely, as illustrated by the following passage from Russell (1918–1919):

> Particulars have this peculiarity, among the sort of objects that you have to take account of in an inventory of the world, *that each of them stands entirely alone and is completely self-subsistent*. It has the sort of self-subsistence that used to belong to substance, except that it usually only persists through a very short time. That is to say, *each particular that there is in the world does not in any way logically depend upon any other particular. Each one might happen to be the whole universe; it is merely an empirical fact that this is not the case.* There is no reason why you should not have a universe consisting of one particular and nothing else. That is a peculiarity of a particular. In the same way, in order to understand a name for a particular, the one thing necessary is to be acquainted with that particular. When you are acquainted with that particular, you have a full, adequate, and complete understanding of the name, and no further information is required. (p. 63, my emphasis)

In speaking of "particulars," Russell means his ultimate metaphysical simples—the momentary sense data named by logically proper names in the logically perfect language he imagines emerging at the end of analysis. Once this stage is reached, our description of the world will include names for all such particulars plus claims (quantified and otherwise) to the effect that they have various simple (unanalyzable) properties and stand in various such relations to one another. The point to notice about the passage is the way in which the *logical* independence of metaphysical simples is linked with their *modal or metaphysical* independence. Does Russell think that each simple concrete particular is *logically* independent of all others because each *could* exist in splendid isolation from all others, or is it the other way around? Although he may not have clearly distinguished the two, it is the latter that is most noteworthy here. What reason is there to think that each such particular could exist all by itself? Let n be a logically proper name of a concrete particular o. Since ⌜~ ∀x x = n⌝ isn't a logical truth, it is *logically possible* for o to be the only existing concrete particular. The implicit (unargued) suggestion is that all logical possibility is modal or metaphysical possibility—in which case, it will follow that it is *metaphysically possible* for o to be the only existing concrete particular. Since the converse inference is likely to seem even more plausible, I suspect that Russell implicitly takes modal necessity/possibility and logical necessity/possibility to coincide (despite the evident implausibility of supposing that a single sense datum, for example a single tactile sensation of hardness, could have existed by itself in the universe.)[22]

[22] Not that he explicitly says so. On the contrary, his use of normally modal terms like "possible" and "necessary" is decidedly idiosyncratic. However, it is hard for anyone to get along without sometimes implicitly invoking metaphysically modal notions—as he does in using counterfactual

Next notice his implicit assertion that for any concrete particular o, the claim that *o isn't the only existing thing* is "empirical," and hence *can be known only aposteriori*. Why would that be? Well, if all apriori truths are logically necessary, then the fact that the claim expressed by ⌜~ ∀x x = n⌝ (where n names o) isn't logically necessary shows that it can't be known apriori. For one who shares Russell's grand vision of analysis, this reduction of epistemic modality to logical modality is well-nigh irresistible. It is a central aim of logical atomism to replace unanalyzed terms, predicates, and sentences/propositions—which may stand in conceptual relations to one another—with logically proper names, simple unanalyzable predicates, and fully analyzed sentences/propositions. When this aim is achieved, the conceptual properties of, and relations holding among, unanalyzed expressions and sentences/propositions are traced to genuinely logical properties of, and relations holding among, fully analyzed sentences/propositions of one's logically perfect underlying language. To take a simple example, it is knowable apriori that all squares are rectangles because the unanalyzed sentence *all squares are rectangles* is reduced, on analysis, to the fully analyzed logical form *all rectangles with equal sides are rectangles*, which is *logically necessary*—and hence knowable apriori. The idea that analyses with similar results can be carried through whenever we encounter conceptual dependencies in unanalyzed language is a driving force behind logical atomism. In this way, Russell is led down a path that makes a reduction of epistemic and metaphysical modalities to logical ones seem plausible.

7. Tractarian Language and the Limits of Intelligibility

When one moves to *The Tractatus*—Wittgenstein (1922)—one finds a superficially similar system of logical atomism that is put to a remarkably different purpose. Like Russell, Wittgenstein posits a multiplicity of metaphysical simples denoted by logically proper names of an imagined logically perfect language. However, his conception of the aim of analysis was not epistemic. Whereas Russell's aim was to find the right substratum for explaining all ordinary and scientific knowledge, Wittgenstein's was to articulate a parallel between language and reality that would make sense of his thesis that metaphysical necessity and epistemological apriority are logical necessity and nothing more.[23] On his picture, metaphysical simples are eternal, unchanging bare particulars with no intrinsic "material" properties of their own, but with the possibility of combining with other simples to form atomic facts.[24] For every possible atomic fact, there is an atomic sentence of the ideal language that would be made true were the fact actually to exist.[25] Just as every atomic sentence (which specifies various simples as standing in one or another relation) is logically independent of every other such sentence, so every possible atomic fact is metaphysically

conditionals in his informal discussion of the reduction in Russell (1914), and as he does in Russell (1918–1919) in discussing possible states that the universe could be, or have been, in.

[23] Chapter 1 of Fogelin (1987) presents a clear and insightful sketch of the essential elements of tractarian metaphysics.

[24] *Tractatus* 2.01–2.033.

[25] *Tractatus* 3.2, 3.203, 3.21, 4.06, 4.1, 4.2, 4.21, 4.22, 4.221.

independent of any other.[26] Just as every assignment of truth values to atomic sentences is logically possible, so every corresponding combination of possible atomic facts is a (complete) genuinely possible way the world could have been. In keeping with this, Wittgenstein held that all meaningful sentences are truth functions of atomic sentences, which is itself quite remarkable despite the fact the he had a rather rich conception of truth functionality.[27] Thus, a specification of the atomic facts making up a possible world-state determines the truth value of every meaningful sentence at that world-state.[28] The "tautologies," which are made true by every assignment of truth values to atomic sentences, are true at every possible world-state; the logical contradictions, which are not made true by any assignment, are true at no possible world-states; and all other meaningful sentences are true at some world-states, but not others.[29] To understand a sentence is to know the possible world-states at which it is true.

Since all possible facts are atomic facts, there are no necessary facts for "tautologies" to state. Since they are true at all possible world-states, understanding them and knowing them to be true doesn't give one any information about the actual world-state—the way things actually are—that distinguishes it from any other merely possible ways things might be. This leads Wittgenstein to claim that "tautologies" are empty: they don't *say* anything.[30] They are simply the result of having a symbol system that includes truth functional operators. So, Wittgenstein thought, whether or not something is a tautology should be determined (and in principle be discoverable) by (examining) its form alone.[31] Thus we have (i)–(iii).[32]

(i) All necessity is linguistic necessity, in that it is the result of our system of representing the world, rather than the world itself. There are sentences that are necessarily true, but there are no necessary facts that correspond to them. These sentences tell us nothing about the world, rather, their necessity is due to the meanings of words (and therefore is knowable apriori).
(ii) All linguistic necessity is logical necessity.
(iii) Logical necessity is determinable by form alone.

With these doctrines in place, it was a short step to the Tractarian test for intelligibility. According to the *Tractatus*, every meaningful statement S falls into one or the other of two categories: either (i) S is contingent (true at some possible world-states and false at others), in which case S is both a truth function of atomic propositions and something that can be known to be true or false only by empirical investigation, or (ii) S is a tautology or contradiction that can be known to be such by purely formal calculations. The paradigmatic cases of meaningful sentences are those in the first category. The sentences in the second category are included as meaningful because they are the inevitable product of the rules governing the logical vocabulary used in constructing sentences of the first category. For Wittgenstein, tautologies and contradictions don't state anything, or give any information

[26] *Tractatus* 2.061, 2.062, 4.211, 5.134.
[27] *Tractatus* 5, 5.2521, 5.2522, 5.254–5.32, 5.501, 5.52–5.523, 5.54, 6, 6.01. For detailed discussion of the account of truth function operations in the *Tractatus,* see chapter 10 of Soames (2003).
[28] *Tractatus* 2.04, 4.26.
[29] *Tractatus* 4.06, 4.46.
[30] *Tractatus* 4.461–4.462.
[31] *Tractatus* 5.551, 6.126.
[32] *Tractatus* 6.1, 6.31, 6.37, 6.3751.

about the world. But their truth or falsity can be calculated, and understanding them reveals something about our symbol system. Thus, they can be regarded as meaningful in an extended sense.

8. The Rise of Logical Empiricism

It was this doctrine about the limits of meaning/intelligibility, more than anything else, that captured the attention of analytic philosophers like Carnap in the 1920s and 1930s. Wittgenstein's route to the doctrine—via an utterly fantastic metaphysics and a completely unrealistic conception of an ideal, logically perfect language, somehow underlying our ordinary thought and talk—were, by and large, treated as baggage that could be dispensed with, so long as the emerging view of philosophy as nothing more than the logical analysis of language—scientific, mathematical, and/or ordinary—could be preserved and extended.

Indeed, it was this uncompromising view of philosophy that emerged from the *Tractatus* which had the strongest, the most immediate, and the most lasting impact. Just as, according to Wittgenstein, the most fundamental ethical claims are neither tautologies nor contingent statements about empirically knowable facts, so philosophical claims are, in general, neither tautological nor contingent statements about empirical facts. Like ethical sentences, they are non-sense. Hence, there are no meaningful philosophical sentences; there are no genuine philosophical questions; and there are no philosophical problems for philosophers to solve. It is not that philosophical problems are so difficult that we can never be sure we have discovered the truth about them. There is no such thing as the truth about them, because there are no philosophical problems.[33]

This view, dramatically expressed in the *Tractatus*, was taken over in less flamboyant form by Carnap and other logical positivists by the early 1930s.[34] The key *Tractarian* inheritance was the distinction between analytic and synthetic sentences, the former being empty of empirical content and knowable apriori on the basis of understanding their meanings alone, and the latter being contingent, knowable only aposteriori, and subject to one or another version of the verifiability criterion of meaning. There was, of course, also a substantial remnant of Russell in the appeal to sense experience as the touchstone of empirical meaning, which in turn led to a resurgence of phenomenalism in some quarters.[35] Despite the official exclusion of metaphysics, ethics, and other forms of traditional philosophical speculation, there was, of course, plenty of work for Carnap and his band of scientifically minded philosophers to do in attempting to work out precise and detailed formulations of what was to prove to be a very difficult view to sustain. It was at this stage that logical and linguistic analysis was officially proclaimed to be the essence of philosophy, rather than merely a powerful tool to be employed in the service of more or less traditional philosophical ends. Thus it is in Carnap (1934) that we find the bold philosophical manifesto,

[33] *Tractatus* 4.11, 4.111, 4.112, 6.42, 6.5–6.521, 6.53, 6.54, 7.

[34] For illuminating early statements of the positivist view, and its debt to Wittgenstein, see, Schlick (1930–1931), Carnap (1932), and the introduction to Ayer (1959).

[35] See, for example, Carnap (1928), Ayer, (1936), Ayer (1940), and Schlick(1934).

"Philosophy is to be replaced by the logic of science—that is to say by the logical analysis of concepts and sentences of the sciences" (p. 292 of the English translation).

9. Assessment

Looking back at the first 55 years of analytic philosophy, from Frege's *Begriffsscrift* to Carnap's *The Logical Structure of Language*, one finds enormous philosophical progress, especially in philosophical logic, the philosophy of mathematics, and the philosophy of language, but few enduring positive lessons about philosophical methodology, and the role of analysis in philosophy. Despite the enormous advance of logicist philosophy of mathematics over what had preceded it, attempts to use the logicist conception of analysis as an excuse to rewrite the content of what is pretheoretically known in other areas to fit largely unexamined preconceptions about knowledge did not meet with very much success. Nor did the linguistic turn in philosophy—based as it was on an unfortunate conflation of the logical, linguistic, epistemic, and metaphysical modalities—provide a solid and broad-based foundation for linguistic and logical analysis. By contrast, Moore's reminder that ordinary pretheoretic convictions have a useful (though limited and fallible) role to play in evaluating philosophical theories was, though modest (when properly understood), of lasting value.

Other than that, my favorite statement of philosophical methodology from the period was almost a throwaway comment by Russell in lecture 8 of Russell (1918–1919).

> I believe the only difference between science and philosophy is that science is what you more or less know and philosophy is what you do not know. Philosophy is that part of science which at present people choose to have opinions about, but which they have no knowledge about. Therefore every advance in knowledge robs philosophy of some problems which formerly it had, and if there is any truth, if there is any value in the kind of procedure of mathematical logic, it will follow that a number of problems which had belonged to philosophy will have ceased to belong to philosophy and will belong to science. (p. 154)

The point, as I would summarize it, is that philosophy is the way we approach problems that are presently too elusive to be investigated scientifically. The goal is to frame questions, explore possible solutions, and forge conceptual tools needed to advance to a more definitive stage of investigation. If one looks back at the philosophers in the first 60 years of the analytic tradition, it is impossible not to be impressed with their seminal contributions to the development of modern symbolic logic in all its present richness, to the mathematical theory of computation, as well as to later advances in cognitive science and computational theories of mind, including the formulation of productive frameworks for investigating the semantics of natural human languages (to which the contributions of Frege and Russell can hardly be overestimated). These are, of course, only a few of the most obvious achievements of many ongoing lines of philosophical investigation initiated during the founding period of analytic philosophy. To my mind, they are far more important than the manifest philosophical failures that seem all too obvious when we focus narrowly on the shortcomings of individual philosophical projects and systems.

References

Ayer, A. J. (1936), *Language, Truth, and Logic,* London: Gollancz, second edition 1946.

Ayer, A. J. (1940), *The Foundations of Empirical Knowledge*, London: Macmillan.

Ayer, A. J. (1959), ed., *Logical Positivism*, New York: The Free Press.

Boolos, George (1994), "The Advantages of Honest Toil over Theft," in Alexander George, ed., *Mathematics and Mind,* Oxford: Oxford University Press; reprinted in Boolos (1998), *Logic, Logic, Logic,* Cambridge, Harvard University Press.

Burgess, John (2005), *Fixing Frege,* Princeton: Princeton University Press.

Carnap, Rudolf (1928), *Der Logische Aufbau Der Welt,* Berlin-Schlachtensee: Weltkreis-Verlag; translated and reprinted *The Logical Structure of the World*, Berkeley and Los Angeles: University of California Press, 1969.

Carnap, Rudolf (1932), "Uberwindung der Metaphysik durch Logische Analyse der Sprache," Erkenntnis, II, translated and reprinted, "The Elimination of Metaphysics Through the Analysis of Language," in Ayer (1959).

Carnap, Rudolf (1934), *Logiische Syntax Der Sprache*, Schriften zur wissenschaftlichen Weltauffassung, hrsg. von Philipp Frank und Moritz Schlick, Bd. 8, Wein: Verlag von Julius Springer; translated 1937, *The Logical Syntax of Language*, London: Kegan Paul.

Fogelin, Robert (1987), *Wittgenstein*, London and New York, Routledge.

Frege, Gottlob (1879), *Begriffsscrift*, Halle, a. S., 1879; sections 1–82; selections trans. by Michael Beaney, Beaney (1997), *The Frege Reader*, Oxford: Blackwell, 47–78.

Frege, Gottlob (1884), *Die Grundlagen Der Arithmetik*, Breslau: Verlag von Wilhelm Koebner 1884; trans. By J. L. Austin, *The Foundations of Arithmetic,* Oxford: Blackwell, 1950.

Frege, Gottlob (1892), "Uber Sinn and Bedeutung," *Zeitschrift fur Philosophie und Philosophische Kritik*, C, 1892; trans. "On Sense and Reference," by Max Black, in Geach and Black (1970), *Translations From the Writings of Gottlob Frege*, Oxford: Blackwell, 56–78.

Frege, Gottlob (1893), Vol. 1, *Grundgesetze der Arithmetik,* Jena; trans. as Vol. 1, *The Basic Laws of Arithmetic*, by M. Furth (ed.), Berkeley and Los Angeles: University of California Press, 1964; selections trans. Michael Beaney in Beaney (1997), 194–223.

Frege, Gottlob (1903), Vol. 2, *Grundgesetze der Arithmetik,* Jena; trans. as *The Basic Laws of Arithmetic*, by M. Furth (ed.), Berkeley and Los Angeles: University of California Press, 1964; selections trans. Michael Beaney in Beaney (1997), 258–89.

Gödel, Kurt (1944), "Russell's Mathematical Logic," in Schilpp (1944), 125–53.

Hodes, Harold (2012), "Why Ramify," unpublished manuscript.

Klement, Kevin (2004), "Putting Form Before Function: Logical Grammar in Frege, Russell, and Wittgenstein," *Philosophers Imprint*, 4/2, 1–47.

Landini, Greg (1998), *Russell's Hidden Substitutional Theory,* Oxford: Oxford University Press.

McTaggart, J. M. E. (1921, 1927) *The Nature of Existence*, two volumes, Cambridge: Cambridge University Press.

Moore, G. E. (1909), "Hume's Philosophy," *The New Quarterly*, November 1909, 545–65; reprinted in Moore (1922), *Philosophical Studies*, London: Routledge and Kegan Paul.

Moore, G. E. (1910–1911), *Some Main Problems of Philosophy*, 1953, New York: Collier.

Moore, G. E. (1925), "A Defense of Common Sense," in J. H. Muirhead, ed., *Contemporary British Philosophy*, 2nd Series, New York: Macmillan; reprinted in G. E. Moore (1958), 32–59.

Moore, G. E. (1939), "Proof of an External World," *Proceedings of the British Academy*, Vol. XXV; reprinted in Moore (1958).

Moore, G. E. (1958), *Philosophical Papers*, London, George Allen and Unwin.
Pryor, James (2000), "The Skeptic and the Dogmatist," *Nous*, 34, 517–49.
Pryor, James (2004), "What's Wrong with Moore's Argument," *Philosophical Issues,* 14, 349–77.
Russell, Bertrand (1905), "On Denoting," *Mind,* 14, 479–93.
Russell, Bertrand (1907), "The Regressive Method of Discovering the Premises of Mathematics," published posthumously in Bertrand Russell, *Essays in Analysis*, edited by Douglas Lackey, New York: George Braziller, 1973, 272–83.
Russell, Bertrand (1912), *The Problems of Philosophy*, London: Williams and Norgate; New York: Henry Holt and Company; Repr. New York and Oxford: Oxford University Press, 1997.
Russell, Bertrand (1914), *Our Knowledge of the External World*, Chicago and London: Open Court.
Russell, Bertrand (1918–1919), "The Philosophy of Logical Atomism," *Monist,* 5, no. 28, 495–527, continued in *Monist*, 5, no. 29, 32–63, 190–222, 345–80; reprinted in *The Philosophy of Logical Atomism*, Peru Illinois: Open Court Publishing, with an introduction by David Pears, 1985. Textual citations are to the 1985 reprinting.
Russell, Bertrand (1919), *Introduction to Mathematical Philosophy*, London: George Allen and Unwin; reprinted in 1993 by Dover.
Russell, Bertrand (1940), *An Inquiry into Meaning and Truth*, London: Unwin.
Russell, Bertrand (1959), *My Philosophical Development*, London: Routledge.
Russell, Bertrand and Alfred North Whitehead (1910), *Principia Mathematica*, Vol. 1, Cambridge: Cambridge University Press; page numbers of citations in the text are to a later edition, *Principia Mathematica to *56,* (1973), Cambridge: Cambridge University Press.
Schilpp, P. A. (ed.) (1944), *The Philosophy of Bertrand Russell* (The Library of the Living Philosophers), La Salle, Ill: Open Court.
Schlick, Moritz (1930–1931), "Die Wende Der Philosophy," *Erkenntnis*, I, translated and reprinted, "The Turning Point in Philosophy," in Ayer (1959).
Schlick, Moritz (1934), "Uber das Fundament der Erkenntnis," *Erkenntnis* IV, translated and reprinted, "The Foundation of Knowledge," in Ayer (1959),
Soames, Scott (2003), *Philosophical Analysis in the Twentieth Century*, Vol. 1, Princeton and Oxford: Princeton University Press.
Soames, Scott (2014), *The Analytic Tradition*, Volume 1, Princeton: Princeton University Press.
Wittgenstein (1922), *Tractatus Logico-Philosophicus,* English translation by C. K. Ogden, London: Routledge.

CHAPTER 4

NINETEENTH-CENTURY AND EARLY TWENTIETH-CENTURY POST-KANTIAN PHILOSOPHY

PAUL FRANKS

1. Introduction

The Post-Kantian tradition had dominated European and North American philosophy for much of the previous century when both analytic philosophy and phenomenology emerged in the early 1900s. I will divide this tradition into three moments: Kantianism, Post-Kantian Idealism, and Neo-Kantianism. I will focus mainly on the original German versions of these three moments, each of which was taken up in more or less innovative ways by philosophers working in other languages. My concern is to explicate the distinctive methods of a tradition that has never entirely disappeared and that is now acknowledged once again as the source of contemporary insights. Indeed, it may be argued that both analytic philosophy and phenomenology, or at least some versions of them, are descendants of this tradition, rather than opponents.

Philosophical methods should be understood and assessed in relation to the problems that they address. Accordingly, I will characterize two problematics that constitute the tradition in question: *naturalist scepticism* and *historicist nihilism*. The first concerned sceptical worries about reason emerging from the attempt to extend the methods of natural science to the study of human beings. In response, Kant pioneered the project of *a critical and transcendental analysis of reason*. At stake was the question whether metaphysics—understood as the human being's attempt, as a rational being, to grasp the intelligible structure of reality—was possible at all and, if so, under what conditions. I will distinguish four methods developed by Kant, each with an assigned role.

No less important was a second problematic that emerged at the very outset of Kant's reception: the problematic of *nihilism*. By the late nineteenth century—thanks to the development of distinctive non-naturalist methods of the *Geisteswissenschaften* or human sciences—nihilism was seen as a threat arising primarily from historicism, understood by

analogy with naturalism as an attempt to make exclusive use of the methods of historical science for the study of human beings and our supposed rationality.

In order to overcome obstacles confronting Kant's original versions of his methods, which I will characterize below, his successors engaged in more-or-less radical revisions of the aforementioned methods. However, in order to minimize presuppositions, the Idealist approach assumed what some considered an inappropriately superior attitude towards the special sciences. Some called for a "return to Kant" that naturalized Kantianism in terms of the developing science of empirical psychology. Others—who came to be known as Marburg and Baden (or South West) Neo-Kantians sought to anchor revisions of Kantian methods in the actuality of the natural and human sciences, in distinction from both the speculative excesses of the Idealists and the psychologistic errors of the naturalist Neo-Kantians.

2. The Success of the Natural Sciences and the Threat of Naturalist Scepticism

If early modern philosophy was largely concerned with the founding of a mathematical science of nature, then late modern philosophy was occupied with the consequences of this project's success and the prospects for its extension to new phenomena, above all to human beings. Hume aspired to be "the Newton of the human sciences" and Kant was the first in Germany to offer lectures on "anthropology".

However, Kant realized that the new science of human nature threatened to undermine the very possibility of another science that he wanted to set on a sure footing: metaphysics, understood as the human being's attempt, as a rational being, to grasp the intelligible structure of reality. Hume's sceptical argument that there could be no ultimate justification for the principle grounding our causal judgements could be iterated for any metaphysical principles. For such principles expressed necessary truths knowable a priori. But they could not be known through general logic alone, since their falsity could be conceived without contradiction. At the same time, they could not be known on the basis of sense perception, which could ground only a posteriori judgements.

This argument had three disturbing upshots. First, we human beings were not as rational as we self-flatteringly pretend. Reason—which is to say, rational inference—played only a minor role in belief-formation and action for these inferences were founded on principles that could not be rationally justified. In fact, these principles expressed, not fundamental truths about reality grasped by humans qua rational beings, but mere *habits*. Just as Newton's physics gave an economical and predictive summation of the regular movements of terrestrial and celestial bodies, organized under a small number of generalizations, so the new human science would reduce the regularities of human belief-formation and action to a small number of generalizations concerning human nature. Second, there was no hope whatsoever for a science of metaphysics comprising knowledge of reality's fundamental character, whether of structural features of sensible events and objects, such as causality and substantiality, or of supersensibles such as God, the soul, and the deep structure of the world. Third, Kant thought that Hume's account required a significant revision of

modern natural science as well. For Newton had understood himself, not as generalizing about observed regularities, but as deducing exceptionless laws of nature.

Underlying these Humean arguments was the assumption of *methodological naturalism*: the idea that *only* the methods of the natural sciences could be used for the acquisition of knowledge, perhaps even for rational belief formation. If one were unhappy with Hume's arguments and their implications, then one could question methodological naturalism itself, developing non-naturalistic methods for the investigation of the human. One could also question Hume's characterization of natural scientific methods. As I understand him, Kant questioned both.

3. The Project of a Critical and Transcendental Analysis of Reason

In response to the threat of naturalistic scepticism, Kant developed a project that may be described as a critical and transcendental analysis of the rational faculty. The rational faculty is the capacity not only to make valid inferences, but also to arrive at *principles* of explanation and justification, and to organize judgements in accordance *systematically*, or in accordance with these principles. Second, as a *faculty*, it pertains only to *finite* rational beings, since an infinite rational being would be fully actualized.

Kant assumed that a faculty has a *form*—here a complex, multi-layered form—as well as a *matter* and an *end,* or an order of ends. The analysis of a faculty was an attempt to determine its proper form, matter, and end. The analysis was *critical* insofar as, equipped with this determination, one would also be able to criticize the use of the faculty for *improper* ends. If it turned out that attainment of the faculty's end was impossible, we humans would be left in a state of perpetual frustration. The analysis was *transcendental* insofar as it showed how the faculty in question could be deployed for the acquisition of a priori knowledge of things as they appear to human beings. In Kant's view, his analysis was also critical because it showed that the faculty of reason could not be used for the acquisition of a priori knowledge of things in themselves. But it is possible to conceive of a faculty analysis that would be critical without being transcendental, ruling out a priori knowledge altogether.

Famously, Kant responded to naturalistic scepticism by formulating what he called the general problem of pure reason: how are synthetic a priori judgements possible? It is important to note, first, that this is a "how possible" question of a specific kind.[1] It is not a question about whether a kind of judgement is in fact actual. Kant assumed the actuality of synthetic a priori judgements as the default understanding that we have of our own cognitive practices, and he would have been prepared to revise this understanding only in the face of abject failure. Nor is the question a request for a mechanism, assuming that the kind of judgement in question is possible, but asking for a clarification of something unknown. Rather, the Kantian "how possible" question expresses a mind-boggle: a perplexity as to how something that we have compelling reasons to regard as actual can be *so much as possible*, when there are apparently compelling reasons to regard it as *impossible*.[2]

[1] See Cassam (2007), 1–10. [2] See Conant (2004).

Conceptual analysis, which Kant understood as *decomposition* into criteria of application, had only subsidiary status. Kant did not imagine the *transformational analysis* pioneered by Frege and Russell, in which deep structure is revealed through translation into a perspicuous notation associated with post-Kantian developments in logic.[3] Such analysis is not obviously incompatible with Kant's project, and would render it redundant only by showing metaphysical principles to be analytic in the sense of following from general logic alone. Even mathematical judgements, however, are generally thought to have been shown to follow from the synthesis of general logic with something else, such as set theory, and the role of intuition in such a synthesis remains debated. In the analytic tradition, rational faculty analysis has been reformulated as what Strawson calls *connective* analysis: an attempt to elucidate "an elaborate network, a system, of connected items, concepts, such that the function of each item, each concept, could, from the philosophical point of view, be properly understood only by grasping its connections with the others, its place in the system—perhaps better still, the picture of a set of interlocking systems of such a kind."[4]

The Post-Kantian Idealists, along with the Baden and Marburg Neo-Kantians, inherited Kant's project of faculty analysis. However, the Idealists were concerned that focus on the human *faculty* of reason encouraged the illusion that reason is subjective, while the Neo-Kantians worried about the suggestion that the proper method of philosophy was psychological. It would be better to say, then, that they were engaged in the critical and transcendental analysis of *reason itself*.

4. Four Methods

Kant intended, by means of this project, to pre-empt the three disturbing upshots of Hume's naturalism. Instead of opposing his methodological non-naturalism to the success of the natural sciences, however, Kant thought that he could show how this success was possible.

In what follows, I will characterize four methods of the analysis of reason from the vantage point attained in Kant's third critique. Kant distinguished three *domains* for the deployment of the faculty of reason: theoretical cognition, in which the end is a systematic organization, under principles of explanation, of cognitions that the world is thus-and-so; practical cognition, in which the end is a systematic organization of cognitions, under principles of justification, that the good demands such-and-such; and teleological judgement, in which the end is a systematic organization of judgements of purposiveness. He also distinguished four *subfaculties* of reason: *sensibility*, the capacity to receive the non-conceptual matter given to reason, whether intuitions in the case of theoretical cognition, or inclinations in the case of practical cognition; *reflective judgement*, the capacity to form concepts that enable communication about given particulars; *determining judgement*, the capacity to cognize particulars by subsuming them under formed concepts, and to make immediate inferences from judgements to judgements; and *reason proper*, the capacity to organize cognitions under unconditioned principles and within a systematic order of ends.

[3] See Beaney (2007), 1–30. [4] Strawson (1992), 19.

Within each domain, all four sub-faculties are operative, but in relevantly different configurations. Each sub-faculty has both a form and an end, and each of the three higher sub-faculties receives its matter from a lower sub-faculty or sub-faculties.[5]

Each critique focuses primarily on one of the three domains.[6] The *Critique of Pure Reason* is principally concerned with theoretical cognition; the *Critique of Practical Reason* with practical cognition; and the *Critique of the Power of Judgment* with teleological judgement.

I will distinguish four methods corresponding to the four sub-faculties. The relation of these methods to paradigmatic sciences or judgemental practices has been a principal issue occupying Kant's successors and interpreters. On the one hand, intimate connection to a paradigm shows Kant's laudable attention to the details of contemporaneous sciences. On the other, significant scientific developments since Kant's time would seem to require significant philosophical revisions, perhaps even, if naturalism and historicism are no longer troubling, abandonment of the project. On yet a third hand, so to speak, Kant assumed a continuity between everyday cognition and science that seems less plausible in light of later scientific developments. Some have therefore sought to construe Kant's project in terms of everyday deployments of reason, which presumably change far more slowly than the sciences, if they change at all.[7]

4.1 Transcendental Aesthetic

4.1.1 Metaphysical Exposition

A *transcendental aesthetic* seeks to establish the form and end of a sub-faculty of sensibility. Theoretical sensibility intuits the given, while practical sensibility expresses itself in inclinations to act. The a priori form of sensibility is established in what Kant—somewhat confusingly—called a *metaphysical exposition*.[8] In the practical case, this form would be the theme of what is now called rational choice theory. For Kant it is the natural meta-inclination to happiness, structured by the principle that one should adopt the necessary means for the implementation of one's maxims, or else repudiate the maxim.[9] In the theoretical case, more important for my present purpose, Kant sought to show that the form of sensibility was non-conceptual in three ways. Taking concepts to be general representations whose application to particulars was mediated by general characteristics, Kant argued that the matter of sensibility was not conceptual but intuitive, where intuitions were immediate encounters with singular items or events. Second, a conceptual form was supposed not only to render intelligible, but also itself to be intelligible. Self-location and object-location within a spatio-temporal world was, for human beings, a genuine and ineliminable way to render things intelligible, but only up to an apparently arbitrary limit. The

[5] Compare Deleuze (1984).

[6] Kant did not originally see things in this way. He expected the *Critique of Pure Reason* to deal with all domains for the deployment of the faculty of reason. At that time, he did not think that practical cognition required a distinct critique, and he did not yet think of teleological judgement as a domain at all.

[7] Contrast Strawson (1966) with Friedman (1992).

[8] Kant (1781/1787), A22–5/B37-40, A30–2/B46–8.

[9] Kant (1788), 5:21–8.

three-dimensionality and chirality of space, and the irreversibility and imperceptibility of time, were brute facts. Humans could establish relations among objects, including their own bodies, through sense perception. But they could not thereby come to know the inner natures or intrinsic properties of embodied things. Third, the form of sensibility was itself known, not as a general set of features pertaining to what was given to sensible intuition, but as given in a priori *intuition*. This last claim was intimately related to Kant's account of mathematics, especially geometry, as involving construction in pure intuition. Through pure intuition, the human being was immediately given an infinite spatio-temporal whole, which could then be delimited, in accordance with Euclidean postulates, in order to prove theorems.

4.1.2 *Transcendental Exposition*

A *transcendental exposition* seeks to show that the ascription to sensibility of an a priori form can alone explain the possibility of a priori cognition.[10] It remains disputed whether the metaphysical and transcendental expositions may be disentangled from Kant's views about mathematics, and whether those views have any plausibility.

4.2 Transcendental Analytic of Concepts

A *transcendental analytic of concepts* seeks to establish the form and end of the sub-faculty of determining judgement. The suggestion, made in passing by Kant, that it has a two-fold structure parallel to that of the transcendental aesthetic, has been found helpful.[11]

4.2.1 *Metaphysical Deduction*

The *metaphysical deduction*, like the metaphysical exposition, attempts to establish the a priori status of the form of a sub-faculty. Unlike the exposition, however, the metaphysical deduction seeks to show the complete intelligibility of the form of determining judgement, a form that is doubly conceptual: it is the form of conceptually articulated judgements, and it is itself to be understood in terms of a systematic organization of meta-concepts or categories.[12]

In theoretical philosophy, Kant assumed the finished state of logic, and he took the table of forms of valid immediate inference as the "clue" for his table of forms of valid explanation. This is not to say that Kant deduced what he called transcendental logic from general logic. On the contrary, he saw general logic as an abstraction from the form of cognitively significant explanation. Thus, for example, the general logical form of modus ponens:

$$\frac{P \rightarrow Q, P}{\therefore Q}$$

[10] Kant (1781/1787), A25/B40–1, A32/B48–9.
[11] Kant (1781/1787), B159.
[12] Kant (1781/1787), A64–83/B89–116.

abstracts from the transcendental logical principle that expresses the form of causal explanation: "Every event has some cause from which it follows according to a law", which entails that, given the law (if an event of type P occurs then an event of type Q necessarily follows), and given an event of type P, an event of type Q must occur. The concept or category of causality thus corresponds to a principle of explanation, from which a principle of inference may be abstracted. However, transcendental logical forms cannot all be read from the general logical forms in this way. Under three headings of Kant's table of categories—quantity, quality and relation of judgments—two forms correspond to general logical forms but a third does not. An independent sense of the end of determining judgement is therefore required.

The universal law formulation of the categorical imperative shows the bond between theoretical and practical cognition most clearly: "So act that the maxim of your will could always hold at the same time as a principle in a giving of universal law."[13] Exceptionless law is the fundamental form of valid explanation in theoretical cognition and of valid justification in practical cognition.

4.2.2 *Transcendental Deduction*

The *transcendental deduction* of the categories that have already been metaphysically deduced is supposed to show that these categories are objectively valid with respect to every possible matter given to the sub-faculty of sensibility. The difficulty lies in Kant's insistence, in the transcendental aesthetic, that sensibility has its own form. It is neither a purveyor of formless matter, nor does it share the form of the sub-faculty of determining judgement. Why, then, should the categorical form of determining judgement necessarily apply to sense-perceptions in-formed by space-time? And why should we be able to act from the motivation of the categorical imperative even when its demands will not maximize our happiness?

On Kant's own account, the hardest problems in his philosophy fall under this heading. His successors sought instead to avoid and pre-empt the problems by means of expanded versions of Kant's other methods.

4.3 Transcendental Analytic of Judgement

The *Transcendental Analytic of Judgement* is concerned with the role of reflecting judgement in the application of the categories to given sensible matter. If the transcendental deduction succeeded in showing *that* the categories apply to all sensible items without exception, this was only in the *implicit* sense that it is necessarily possible to make judgements about these items that presuppose categorical principles. It remained to be seen *how* the categories apply, and how it is possible to reflectively form concepts of sensible item types and features, in order eventually to arrive at explicit categorical judgements.

In theoretical cognition, the transcendental analytic of judgement sought to fill this gap by showing how pure intuitions of space–time yielded *schemata* mediating between

[13] Kant (1788), 5:30.

categories and sense perceptions. For example, the temporally successive character of sensible appearances enabled the formation of causal concepts, at the limit of which lay the formation of concepts figuring in exceptionless causal laws.[14] In practical cognition, the gap could not be filled in this way, since the moral law was not only exceptionless but also unconditioned, and therefore could not be instantiated by the structural features of space and time, which remain conditioned by ineliminable arbitrariness. Kant instead invoked a law that served as a *type* of the moral law by determining a possible world as a whole. Thus, for example, we ask ourselves whether a given maxim is rationally permissible by considering whether a world in which this maxim served as universal law would be coherent and, if so, whether it would be good.

In both these cases, the goal of reflection is determining judgement: the subsumption of a particular event or maxim under a universal concept or determinate rule, enabling us to say, not only whether, but why the event occurred, or why the maxim is permissible or impermissible. However, in the *Critique of Judgment*, Kant introduced cases where there can be no such goal.[15] He argued that teleological judgements, including judgements regarding beauty or sublimity, could never arrive at universal concepts or determinate rules. Yet such judgements were not mere expressions of subjective taste about which no rational discussion could occur. Rather, these judgements were grounded in features of human experience—notably the possibility of an intrasubjective harmony among the sub-faculties solicited by the mutual attunement of human sensibility and the sensible world—that could never be fully conceptualized but that could be *exemplified* in communicable ways. Here rational discussion involved, not the formulation of determinate and universal rules, but rather the appeal to and the generation of exemplary, particular cases, along with the endeavour to make communicable to each other the way in which our shared humanity grounded our responses.

Kant's account of reflective concept-formation in both theoretical and practical cognition is extremely suggestive but incomplete. In order to fill in the details, one would have to say more about issues raised in the *Critique of Judgment* concerning the communicability and exemplarity of human responses to particulars. Both the Post-Kantian Idealists and the Baden and Marburg Neo-Kantians contributed to the development of this Kantian project.

4.4 Transcendental Dialectic

The analysis of reason has a critical dimension insofar as it isolates necessary conditions for the proper deployment of the faculty. Transgressions of those conditions would be improper uses of reason. However, Kant thought that that there was also a "*logic of illusion*": a systematically structured set of illusions to which the human being is naturally and unavoidably subject.[16] The way to study this structure was to consider the effects on

[14] Kant (1781/1787), A189/B232–A211/256.
[15] Kant (1790), 10:221–4.
[16] Kant (1781/1787), A293/B249–A309/B366.

the conceptual and judgemental apparatus resulting from the quest to ground cognition in *first principles* or *unconditioned premises*. Just as the clue to transcendental analytic was suggested by the general logic of immediate inference, so the clue to transcendental dialectic was suggested by the general logic of the syllogism. Kant used the term "ideas" for the syllogistic analogues of the categories. He thought that they corresponded, not only to specific illusions, but also to the three areas of rationalist special metaphysics. Thus the idea of the unconditioned subject of predication—the infinitely active "I" or the immortal soul—was both the meta-concept of the paralogism or apparently valid categorical syllogism, and at the same time the founding notion of the subfield of rational psychology. The idea of the unconditioned series or whole of explananda—the infinitely intelligible world—was both the meta-concept of the antinomy of hypothetical syllogisms and at the same time the founding notion of the subfield of rational cosmology. Finally, the idea of the unconditioned and systematic interrelatedness of every possible being—the determinacy of every concept of a being in the divine mind, or the absolute community of the real in God—was both the meta-concept of the apparently valid disjunctive syllogism, and at the same time the founding notion of the subfield of rational theology.

Neither rational psychology nor rational theology, on Kant's view, led to contradictions. Although their invalid syllogisms did not lead to cognition, as rationalist metaphysicians had hoped, their founding ideas could be used as *regulative ideas*, guiding the pursuit of cognition towards an unattainable yet still valuable goal. Rational cosmology was a special case, because it led to contradictions or, more precisely, antinomies, pitting against one another syllogisms with opposing conclusions yet of equal force. In some cases, one could avoid sceptical suspension of judgement by abandoning, with transcendental idealism's help, an assumption underlying both opposing syllogisms. However, in what would prove the most interesting case for Kant's successors—the Third Antinomy, in which an argument for causal determinism was opposed by an equally cogent argument for causal spontaneity—Kant used transcendental idealism to retain both syllogisms while mediating their opposition.

In the *Critique of Pure Reason*, Kant merely assigned determinism to the knowable and phenomenal realm, while assigning spontaneity to the merely thinkable and noumenal realm.[17] This did little to make agency intelligible, as Kant recognized. In the second *Critique*, after several attempts at other transitions to moral philosophy, Kant introduced what he called "the fact of reason": an actuality whereby the efficacy of my causal spontaneity within the phenomenal realm was manifest.[18] While Kant's successors were dissatisfied by the details of his presentation, some were inspired by the thought that an antinomy could be resolved through the formation of a concept or idea uniting within itself what would otherwise be contradictory elements. This suggested the possibility that Transcendental Dialectic could be not only a critical logic of illusion, but also a transcendental logic of truth.

[17] Kant (1781/1787), A532/B560–A558/B586. [18] Kant (1788), 5:30–2.

5. Transcendental Arguments and their Problems

What analytic philosophers have come to call a *transcendental argument* is an instance of a format in which Kant presented one of his results. The refutation of (empirical or material) idealism[19] begins with an assumption presupposed by a familiar "Cartesian" variety of scepticism about the external world: that I have immediate awareness of my mental representations as succeeding one another in time. It proceeds to argue that awareness of succession requires equally immediate awareness of something spatial and relatively persistent as the substratum of change. But empirical idealism is the view that awareness of something spatial must be *inferred* from immediate awareness of mental representations. So, if Kant's argument is sound, then empirical idealism is refuted. Moreover, Kant's argument undermines the sort of scepticism that doubts the possibility of a valid inference from mental representations to the external world.

Similarly, transcendental arguments are sometimes said to start from a thesis conceded by the sceptic and to advance therefrom to a necessary condition for the possibility of the truth of this thesis that refutes scepticism.[20] However, reflection on Kant's exemplary argument suggests that this is problematic. Kant claimed to refute *empirical idealism*, not *scepticism*. To be sure, his argument was supposed to undermine one sort of scepticism. But it could hardly claim to refute scepticism once and for all. Does the argument I have sketched exclude the possibility that empirical realism—the thesis that we have immediate awareness of spatial items—can *also* be refuted? If such a refutation were also to be found, then we would be confronted with an antinomy, and the rational response, failing all else, would be to suspend judgement—in other words, scepticism. To judge Kantian and post-Kantian methods solely or mainly by their success in refuting scepticism is a mistake. Instead of seeking to refute scepticism, these methods may be seen as endeavouring to disclose what Cavell calls "the truth of scepticism".[21] In other words, scepticism expresses a truth about human reason that need not be expressed in the form of doubt—for example, a truth that may more perspicuously be expressed, from a Kantian standpoint, as the thesis that empirical realism presupposes transcendental idealism.

I note also that the Refutation of Idealism is an *epitome* of Kant's arguments in the analytic of concepts and the analytic of judgement: of his argument in the First Analogy that awareness of temporal succession requires awareness of spatial persistence, which in turn presupposes the Schematism, which presupposes the Transcendental Deduction, which in turn presupposes the Transcendental Aesthetic. If one were to focus on the Refutation of Idealism alone, one would not be able to reconstruct Kant's thinking in any detail. We should not be overly obsessed with the format in which Kant presented his results, as if he could have arrived at these results by organizing his arguments in this way in the first place.

[19] Kant (1781/1787), B274–9.
[20] See Stern (1999), (2000).
[21] Cavell (1972), 107n.; (1979).

5.1 The Uniqueness Objection

Still, the transcendental argument format helps to present some objections to which an advocate of rational faculty analysis will have to respond. Consider, then, an argument presentable in the format, "X is a necessary condition for the possibility of Y", where "X" stands either for some aspect of reason that could not be established naturalistically, or for some aspect of the world's intelligible character that could not be established naturalistically and about which naturalism therefore enables scepticism.

Note, first, that there may be a *disjunction* of distinct and separately sufficient conditions for Y's possibility. Even if X is *one* necessary condition for the possibility of Y, it may be only one of several alternatives. It remains possible that another condition *actually* enables Y. So X has not been established and scepticism remains untouched. Moreover, if another sufficient condition for the possibility of Y is naturalistically acceptable, then naturalism remains untouched too. This is often formulated as a complaint that X has not been shown to be *uniquely* sufficient for the possibility of Y.[22]

One response would be to show that X is the *only* necessary condition for the possibility of Y that meets additional constraints on a philosophically satisfying account. Another is to show that X is the *actual* condition enabling Y by linking X to some well-established cognitive practice, such as mathematics, general logic, natural science, or everyday ethics. A third, characteristic of Strawson's analytic renewal of Kantianism, is to try to conceive Y without X. If the experiment fails, then X must be a necessary condition for Y's possibility.

5.2 The Reality Objection

Another objection is that arguments for the conclusion, "X is a necessary condition for the possibility of Y", can establish at most that, "The presupposition that X is a necessary condition for the possibility of Y". Something further is required for the inference to the *reality* of X. Without this further something, neither scepticism nor naturalism appear to be touched by the argument.[23]

Stroud pointed out that Kant's arguments were untouched by the Reality Objection,[24] for Kant thought that a conclusion could be established about the empirically real only by means of the assumption that empirical reality is constituted by the forms of the human faculty of reason—including the forms of the sub-faculty of sensible intuition, which could not be forms of things as they are in themselves. Transcendental idealism thus provided the further something to establish X as real. As Stroud further argued, many philosophers saw transcendental idealism as compromising rather than supporting empirical realism, or even as incoherent. Furthermore, since analytic philosophers had come to be interested in the necessary conditions for the possibility of Y in cases where "Y" stood for something thematized as meaningful, they were in need of something further that typically took the form of a theory of meaning. But these theories often ruled out the meaningfulness of sceptical doubt directly. Consequently, transcendental arguments were not necessary for the refutation of scepticism.

[22] Körner (1967). [23] Stroud (1967). [24] Stroud (1994)

As I have suggested, however, it is a mistake to see the refutation of scepticism as the principal goal of Kant or his successors. Stroud himself acknowledged that transcendental arguments had another use in contemporary philosophy: to refute the conventionalist view that whether we use one framework of cognitive forms or another is a pragmatic question.[25] In the view I have taken here, the most important goal of the Kantian and post-Kantian tradition has been to counter the scepticism arising from naturalism and the nihilism engendered by historicism, and more generally to resist the downplaying of the role of reason in human life, while still accounting for the success of our cognitive practices, including the natural and human sciences. Both the Post-Kantian Idealists and the Neo-Kantians thought that this project required idealism in some sense, although they disagreed both with Kant and amongst themselves about what idealism amounted to. In any event, idealism did not render transcendental arguments—or the complex strategies they epitomized—redundant.

5.3 Naturalistic Appropriation

It was observed early in Kant's reception that his arguments typically involved a premise unacceptable to Hume. In contemporary terms, transcendental arguments often beg the question against naturalism by assuming a non-naturalistic construal of "Y". For example, Kant's arguments in the *Critique of Pure Reason* concerned the necessary conditions for the possibility of experience, where experience was understood, not as a bombardment of sensory impressions, but rather in a rich cognitive sense as a body of judgements about the world, grounded on sense perceptions.

One response could be to accept the point and to argue for the privileged status of the transcendental argument's premise on the ground that it is identical to—or is closer than the naturalistic account to—the everyday conception of experience (or whatever Y is), or the scientist's self-understanding. But it is disputable what sort and what degree of privilege this would be, as well as whether it is true. It may well occur to someone interested in the state of play between naturalism and transcendentalism to ask, in light of the aforementioned objection, how the naturalist might regard, not only the premise, but also the *necessity* of a transcendental argument.

Maimon asked just this question. Although sympathetic to Kant's transcendental approach, he argued that the naturalist could take the same attitude to transcendental necessity—the necessity with which X conditions Y—as Hume had taken to causal necessity. In other words, the naturalist could construe the conditional first as entirely *internal*—as a dependence relation between capacities, or between a capacity and a presupposition, rather than as leading to any a priori cognition of the empirical world—and secondly the naturalist could argue that this dependence was the result of a *human habit or custom*, not an expression of strict necessity. This would amount, in contemporary terms, to an anti-realist or quasi-realist appropriation of the transcendental analysis of reason. It would be an attempt to accommodate transcendental philosophy's

[25] Stroud (1967), 243.

results while continuing to downplay, in Humean fashion, the role of reason in the world and in our lives.[26]

6. The Rise of the Historical Sciences and the Spectre of Historicist Nihilism

Kant's successors needed to overcome, not only the obstacles confronting any analysis of reason intended as a response to naturalistic scepticism, but also a distinct problematic emerging simultaneously with Kant's philosophy.

Jacobi argued that rationalism led inexorably to Spinozism, which inevitably determined the characters of finite things, not by their intrinsic or positive properties, but by their negative properties—in other words, by their roles within the world as a whole, hence by their relations to other members of the whole.[27] In footnotes, however, Jacobi noted that something similar could be seen in the account of space and time in the Transcendental Aesthetic of Kant's *Critique of Pure Reason*. The form of space–time constituted a singular whole within which any empirical object could be located and related to any other, but only in virtue of their relations, not in virtue of their intrinsic properties. Similarly, the transcendental unity of apperception was similar to Spinoza's substance or so-called God. Each was immanent to the whole that it served to unify, so each lacked any determinate character of its own.

To generalize, the objection is that transcendental methods may arrive at conditions for the possibility of natural science that avoid naturalistic scepticism while entailing a conclusion that is no better: the non-individuality of finite objects and the indeterminacy of the unifying and immanent first principle. Jacobi called this conclusion "nihilism".[28]

The antidote, Jacobi thought, lay in *history*. Anybody who sought an absolute justification of their philosophy would in due time, if they were sufficiently intelligent, consistent, and honest, arrive at something very like Spinozism or, better yet, a synthesis of Spinozism and Kantian idealism of the sort developed by the Post-Kantian Idealists. But somebody sufficiently immersed in their own ethical context would have unshakeable trust in their own norms, and would see no need for justification.

How could moderns attain unwavering conviction, given their awareness of the diversity of cultures and religions, and the inescapability of a philosophically informed discourse about the comparative justifications of distinct lifestyles? Some sought a substitute for ancient näiveté in historiography.[29] The modern historian was developing a new, non-naturalistic method for the scientific study of human beings. By tracing the development of individuals and individual nations from a specifically national viewpoint, the historian would form reader as individuals committed to the norms of a national culture.

This later seemed a miscalculation. What Jacobi called nihilism—the dissolution of individuality, the loss of purposiveness, still more generally the dissipation of significance—and ascribed to non-naturalistic philosophies seeking to explain the possibility of natural

[26] Maimon (1790). Franks (2007). [27] Jacobi (1785/1789). Franks (2005), 170–4.
[28] Jacobi (1799). [29] Humboldt (1822).

science, re-emerged within the historical approach to the human. Larger contexts subsumed individual persons, and the vicissitudes of nations resisted any teleological account. Historians were at least as good at dissolving supposed unities and at debunking putatively meaningful narratives as they were at forming individual citizens.

Pioneering the human sciences at an early stage in their development, Kant seems not to have anticipated this problematic. But his post-Kantian Idealist successors sought to develop alternative approaches to historiography that gradually came into conflict with practitioners of new historical methods. By the Wilhelmine period, in which German philosophy was dominated by the Neo-Kantian schools of Marburg and Baden, historicist nihilism was widely perceived as a problematic distinct from and no less urgent than naturalistic scepticism, which had come to be known as psychologism. These twin problematics of psychologism and historicism, along with the aforementioned problems encountered by Kantian methods, largely determined the options available to both Post-Kantian Idealists and Neo-Kantians.

7. Post-Kantian Idealist Revisions of Transcendental Aesthetic and Transcendental Dialectic

The Post-Kantian Idealists—principally, Reinhold, Fichte, Schelling, and Hegel—were deeply impressed by Jacobi's provocative claim that only a system such as Spinoza's could hope to satisfy reason's most rigorous demands, especially the demand, not merely for a relative justification that ends in an unjustifiable assumption or that leads to an infinite regress, but for an absolute justification that would terminate non-arbitrarily in a first principle. Such a system, Jacobi argued, would have a first principle that was first in the order of explanation or justification, but that was immanent within—and not transcendent of—the whole of the items explained or justified; meanwhile, the items in question would be determinate in virtue of their roles and relations within the whole. Qua member of the whole, each item would have a relative explanation or justification that could be put in naturalist terms. But, qua manifestation of the first principle, each item would have an absolute explanation or justification.[30]

One way to understand the Idealist project is to see it as combining this picture of an adequate philosophical system, which is not Kant's, with two parts of Kant's methodology: the part of the analytic of concepts called the metaphysical deduction and the part of the Dialectic called the transcendental ideal.

It was in his account of the goals of the metaphysical Deduction that Kant sounded closest to the Spinozist conception of systematicity articulated by Jacobi. Kant emphasized that the table of categories had to be complete: "to entirely exhaust the entire field of pure understanding."[31] However, the Idealists found Kant's execution to be flawed. His idea of the whole appeared rooted in the idea of a finite rational being equipped with a sensibility

[30] See Franks (2005), 99–108. [31] Kant (1781/1787), A64–5/B89–90.

whose forms involved ineliminable arbitrariness, not in the idea of absolute reason. His derivation of the categories assumed that general logic was not only a finished science but was also a more or less transparent guide to the deeper structures of transcendental logic. And it gave rise to categories that were utterly distinct in their derivation from the forms of sensibility with which they had to be conjugated in order to guarantee cognition, generating the problem addressed by the transcendental deduction, which the Idealists considered insoluble if left in the terms in which it had been raised.

On the other hand, when the Idealists looked at Kant's idea of absolute reason, they had to contend with the fact that this was for him at once both indispensable and dialectical in a way that gave rise only to criticism, not to transcendental knowledge. This seemed to them to be a premature capitulation to naturalism. They were particularly struck by Kant's argument, in the section of the Dialectic dealing with rational theology, that determinate things should be situated within the whole of possible reality (*omnitudo realitatis*), which was grounded in a most real being (*ens realissimum*).[32] Kant had refined this argument when he was still a rationalist metaphysician. In his critical philosophy, he came to see it as a need of reason giving rise to a mere idea that could play a regulative role for theoretical cognition, and that could be interpreted to some extent as the structure of the postulates of practical faith. His version, however, depended on the Leibnizian assumption that determinate things would have to be individual substances with intrinsic properties (*entia* per se). Such things could not be located within the sensible world, which Kant thought, after deep reflection on Newtonian physics and on the underlying metaphysics of space and time, had to be relational all the way down. So Kant relegated things in themselves to the merely thinkable or intelligible world. Working with an alternative, Spinozistic view of things as determinate only in virtue of their roles and relations to other members of the whole, the Idealists saw no compelling reason for this relegation. They could adopt a "two aspect" account of phenomena and noumena. From the empirical perspective, sensible things were nodes within the dynamic structure of *natura naturata*. From the transcendental perspective, they were manifestations of *natura naturans*, a first principle understandable only by means of a non-naturalist method.

Putting these two thoughts together, we get the following picture. The Idealists sought a transcendental analytic of concepts that would derive the fundamental concepts or forms of intelligibility from the absolute first principle of reason. This principle could be seen as the immanent unifier of the whole constituted by the forms of intelligibility. At the same time, the resulting list of forms would not, in the end, correspond exactly to Kant's list. First, general logic no longer played a privileged role. Second, the derived forms were meant to include space and time, the forms of sensibility, which would pre-empt the transcendental deduction problem altogether.

But how was this ambitious revision of the transcendental analytic of concepts supposed to be carried out? Two principal methods were developed. Each has several variants, and they have sometimes been combined, but here I confine myself to describing their ideal types.

The first method is construction in a priori intuition. This is based on Kant's transcendental aesthetic, which specifies the a priori character of a form of intuition that can itself

[32] Kant (1781/87), A567–83/B595–661.

be known through intuition in mathematics. However, the Idealists sought a radical extension of this method, so that it would allow knowledge of the absolute first principle and derivation of a list of forms that would include those assigned by Kant to sensibility, understanding, and reason.

Fichte pioneered this method. He was attracted by the idea that the objections raised to Reinhold's early version of the Idealist project—the Uniqueness Objection, the Reality Objection, and the threat of Naturalist Appropriation—did not seem to apply to mathematical construction. Could such construction, rather than general logic, provide the clue to the transcendental analytic of concepts?

A constructive method had three requirements. The first was an analogue of the description of a space in Euclidean geometry—that is, the constitution of a whole within which constructions may be carried out. The second was an analogue of the Euclidean postulates, understood as iterable constructive procedures. Third, in order to be genuinely intuitive, both the constituted whole and the constructions within it would have to be ineliminably particular and incapable of representation by general concepts. In addition, of course, the constraints of Spinozistic systematicity would have to be met. Moreover, since the Idealists were keenly aware of Jacobi's argument that this conception of systematicity annihilated individuality in an intolerable way, they had either to show that Jacobi was wrong or else to show that nihilism was palatable.

It occurred early to Fichte that the first requirement could be met insofar as the self-reference presupposed by any representational and conative act whatsoever—which he called self-positing—was a reference to a principle that both constituted the whole "space" of the thinkable and that exhibited a particularity that could not be reduced to general and conceptual terms. The equivalent to Kant's *ens realissimum* would therefore be the self-positing I, positing itself as self-positing, although Fichte did not want to say either that this I was real or that it was a being. Rather, it was the self-active and ideal principle of all reality and all being. Fichte also argued that at least one construction was possible within the logical space constituted by the self-positing I: the construction of the space of the sensible world. Initially, he doubted, however, that any further constructions were possible by means of iterable procedures similar to the tracing and extension of lines and to the generation of a circle around a point in Euclidean geometry.

Later, Fichte thought that he had arrived at such procedures thanks to the Principle of Determinability: "It is only through opposition that it is possible to obtain a specific and clear consciousness of anything whatsoever."[33] The active self-positing or self-intuiting of the I led inexorably to the opposite positing of the not-I or to the positing of the I in a state of repose. This led in turn to the positing or intuiting of the sensible world within which I and not-I can interact, and within which a plurality of I's can interact, and so on. Starting from a first principle for which no further explanation or justification could possibly be demanded, the deduction of forms of intelligibility would proceed until it reached the limit of explanation, where explanation and justification would lose their sense. Fichte emphasized that reason presupposes a ground that it cannot grasp.

There were of course many problems with this ambitious proposal. Even if the details could somehow be worked out, the Reality Problem seemed especially threatening. After

[33] Fichte (1978), 31.

all, Kant's transcendental aesthetic and metaphysical deduction were anchored in the well-established sciences of mathematics and logic, but Fichte had severed the link to logic, and his constructions in intuition were remote analogues of mathematical constructions. Moreover, Kant himself regarded mathematical constructions as constructions of the forms of sensible objects, not as the constructions of realities. These forms were mathematically unimpeachable, but only in the transcendental deduction, through their connection to cognizable objects, were they philosophically justified. Fichte intended to pre-empt the problem of the deduction. But it was arguable that he had instead succeeded only in exacerbating the problem by deducing mere forms—of sensibility, understanding, and reason—whose reality could not be established at all. If the Reality Objection is combined with Jacobi's charge that Spinozistic systems like Fichte's annihilate individuality, then it is hard to see how Fichte's endeavour, even if worked out to a maximal degree, could present anything but chimeras. The self-positing I was not the individual "me" with a first and last name; the spaces of interaction were mere forms of the world, not the concrete world of actual existence and action; and the derived forms of intelligibility were mere spectres, not animating spirits of the living world.

At the same time, the Uniqueness Objection could be combined with Maimon's insistence that systematizing Kant's philosophy could never refute naturalism. The objection was not that some other Idealist could carry out Fichte's project in some other way. It was rather that the naturalist could also carry out a systematic derivation of forms of intelligibility. Spinoza himself sometimes served as the paradigm. Perhaps Jacobi was right that Spinoza's own system annihilated individuality. But could Spinoza's system not be lived by a human being who accepted the annihilation of his or her individuality as an ethical demand? Was it not then a question, in the end, of the system whereby one wanted to think and live? What gave one system an advantage over the other?

We may think of Schelling and Hegel, during their Jena partnership, as developing another version of Fichte's method of construction in intuition in response to these problems. If possible, this version was even more ambitious. In order to pre-empt Naturalist Appropriation, it would be necessary, so the thought went, to show that the derived forms of intelligibility were forms *of nature*. One way to do this was to start, not from "the I", which made the project sound subjectivist and left nature either out of the story or at least until too late in the story, but from a first principle manifest in nature as well as in the human mind. There would be two parallel derivations—a philosophy of nature and a philosophy of spirit—capped by a demonstration that a single absolute principle realized itself in both. If Fichte's approach left the natural sciences to their own devices and therefore seemed overly subjective, this approach purchased objectivity at the price of intervening in the natural sciences to a more or less significant extent. Mathematical construction was jettisoned as the model. But it was replaced by a method of construction in intuition modelled on Goethe's morphology, according to which the scientist, who was closely related to the artist, could "see" an underlying ideal and dynamic structure manifest in a range of variations, for example the original form of plant in each part of the plant, or the archetype of the vertebrate skull. Now the order of derivations traced the order of evolution itself.[34]

[34] See Förster (2012), 250–76.

Once again, there are many difficulties. There is a heavy investment in a particular way of doing natural science. The sort of intuition involved is similar to that of the artist and just as unequally distributed. Does it matter if philosophical constructions are unintelligible to all but a self-professedly gifted few?

At a time when Fichte did not think that there were iterable constructive procedures in philosophy, he developed an alternative method based on Kant's transcendental dialectic instead, or rather on the transition from the dialectic of theoretical cognition to its resolution in the analytic of concepts of practical cognition, where the antinomy between free causality and determined causality is resolved by the actuality of a free causality that manifests itself in the empirically real, which Kant calls "the fact of reason". Fichte saw a way to start the generation and resolution of antinomies non-arbitrarily—in the opposition between the I self-positing itself as self-positing and the not-I required by the Principle of Determinability to make sense of the I—and he saw a way to continue generating and resolving antinomies until he reached the limit of explanation. Dialectic became, not only a critical way to diagnose illusions of reason, but also a transcendental way to attain metaphysical cognition.[35]

It was after developing this antinomic method that Fichte returned to the method of construction in intuition, which he now thought could be implemented with the help of the Principle and by means of iterated constructive procedures. As I have already mentioned, Schelling—joined by Hegel—found Fichte's version of the constructive method too "subjective": too susceptible to the Uniqueness Objection because it could not refute naturalism, too vulnerable to the Reality Objection because it could not demonstrate the instantiation of its deduced forms to natural beings, and too easily appropriated by naturalism. Accordingly, he developed a new version of the constructive method, based not on Kant's conception of geometric construction in intuition, but on Goethe's conception of aesthetic and teleological construction in intuition. When Hegel came to think that Schelling's Goethean method was itself too subjective, because it was inaccessible to all but a few, and too abstract because it failed to show the identity of nature and spirit in any determinate sense, he developed a revised version of the transcendental dialectic.[36] Building on Fichte's insights, Hegel's method consisted in the iterated generation and resolution of tensions. However, there were two differences from Fichte's version. First, unlike Fichte's "subjective" version, Hegel's traced the generation and resolution of tensions through hierarchical levels of nature itself, as well as through the stages of human history. Second, Hegel's dialectic was far more context-sensitive than Fichte's use of the specifically antinomic pattern of thesis and antithesis and deployed many patterns of tension and resolution.[37]

Schelling also, in his later doctrine of potencies, developed a dialectical method. He argued that Hegel's version, which jettisoned the appeal to intuition that could alone prove the reality of dialectically derived forms, was incapable of responding to the Reality Objection. Dialectic without intuition was empty, while intuition without dialectic was blind. At best, Hegelian dialectic was merely negative philosophy, which demanded supplementation by positive philosophy dealing with existence itself.[38] At the same time, both Schelling and Hegel sought to respond to the threat of nihilism emerging from incipient

[35] Fichte (1794). [36] Hegel (1807) and (1812–16). [37] G. Mueller (1958).
[38] Schelling (1856–61), division II, vol. 3, 74–94.

historicism. If dialectical patterns could be traced in history, then history could be seen, not merely as a succession of epochs with their own incommensurable standards of meaning, but as purposive and thus as trans-historically meaningful. For Hegel, the end of history towards which the dialectic moved was an equilibrium in which individual humans could participate in a self-aware, collective life of spirit. The Reality Objection could be dealt with insofar as philosophy self-consciously took up its position in the present moment, a moment when, with the modern realization of religious, natural scientific, political, and economic institutions and structures, the seeds of this equilibrium could be recognized.

Of course, many difficulties must be confronted by anyone who wants to respond to naturalist scepticism and historicist nihilism by means of transcendental dialectic. Is the development of concepts to resolve antinomies and tensions anything more than a heuristic device? On what basis could we say that such a development inheres in reason itself, let alone—as Hegel and Schelling want to say—that it inheres in the rational character of nature and history? Does such a method not rely upon metaphysical presuppositions that are themselves incapable of any justification beyond their fruitfulness in producing systematic endeavours, whose successes, even in their own terms, are impossible to assess from an external perspective?

8. Neo-Kantian Revisions of Transcendental Analytic of Judgement

The call to go "back to Kant" was first sounded by the fiercest critics of Post-Kantian Idealism during its heyday, from 1790 to 1830. Post-Kantian Idealism seemed too removed from the successes of natural science that had fuelled not only naturalism but also Kant's response to naturalistic scepticism. But these critics did not intend a return to Kant's original methods. On the contrary, they thought that *both* Hume's naturalism and Kant's anti-naturalism were methodologically problematic. Hume had been led to scepticism by misunderstanding the methods of natural science, and the non-naturalist methods of Kant and the Post-Kantian Idealists had accordingly been based on a fundamental error. What was needed was a better understanding of naturalist methods for philosophy to emulate.[39]

For these early proto-Neo-Kantians, the methods in question were those of introspective psychology, applied to "the facts of consciousness". However, a new turn was taken with the development of physiological approaches to psychology by the 1840s. Müller had shown by means of extensive experimental data that a single causal stimulus could bring about qualitatively different effects on distinct sense modalities. The representation's character was determined, not by mind-independent objects, but by the character of human sensibility.[40] A new generation of Neo-Kantians concluded that Kant's transcendental idealism had been vindicated under a naturalist interpretation, and called for a return to Kant from the excesses of Post-Kantian Idealist metaphysics as well as from the materialism of what

[39] Fries (1807). [40] Müller (1826), xii–xiii.

they considered to be inferior, anti-Kantian construals of natural science and its underlying philosophy.

Helmholtz, for example, considered, not "the facts of consciousness", but rather "the facts in perception": the facts in *perception* and not consciousness, because the Newtonian method—a careful mix of experimentation and theory formation, aiming at the formulation of exceptionless laws—could be applied to the *interface* between mental and physical: the facts *in* and not merely *of* perception, because perception was sometimes illusory, and common-sense reflection might turn out to be no more reliable when measured against the standard of a theory that undertook to account for both veridical *and* illusory perception.

Psycho-physiological Neo-Kantians interpreted Kantianism in an empiricist way. Idealism was not transcendental but empirical, and closely resembled the material idealism that Kant had sought to refute. Sensibility operated by means of "unconscious inferences" rather than intuitions; and the cognitive faculty's form was to be discerned by detecting invariants in human cognition, not through a transcendental investigation of human *reason*.

Without a metaphysical or transcendental deduction, Helmholtz could conclude only that "the law of sufficient reason is nothing more than the *urge* (*Trieb*) of our understanding to bring all our perceptions under its own control."[41] How did this differ from Hume's conclusion that causal inference was habitual and ineliminable, but not an expression of reason?

Criticizing the Neo-Kantianism of their teachers, the next generation of Neo-Kantians called for a further return to Kant that would answer this Humean question. In their criticisms of what they called *psychologism*, the Baden and Marburg schools followed the lead of Lotze, contributor to both physiology and philosophy.

Lotze distinguished *the eternal validity of ideas*, first thematized by Plato, from *the temporal actuality of beings*. Beginning with Aristotle, metaphysical philosophers, including Post-Kantian Idealists, had reduced being to validity. On the other hand, empiricists, including the psycho-physiological Neo-Kantians, had reduced validity to being. What was needed was a *pure logic of validity*, purified of all being, including psychic being. Particularly emphasized was the need for an account of concept-formation other than the Aristotelian–empiricist model of abstraction. On that model, more general concepts were formed by omitting content. But although this was psychologically intelligible, it could not explain how we form genuinely universal concepts, for example in mathematics, where more general concepts serve as rules from which less general concepts are derivable.[42] Thus the Baden and Marburg schools' main project was to develop *a purified transcendental analytic of judgement*.

What was wrong with psychologism? On the Baden view, the problem was that validity was governed by norms that cannot be grounded in any descriptive science of beings. Philosophy needed to receive data from the natural and historical sciences, but these data had to be criticized and systematized in accordance with philosophy's understanding of "an ideal of the normal human being".[43] This ideal could alone ground norms, and

[41] Helmholtz (1850–67), 3:455.
[42] Lotze (1874), book 3, chapter 2.
[43] Windelband (1884), 1–53.

comprised the ends or regulative ideals of various disciplines. The end of natural science was the determination of absolutely valid law, and the end of historical science was the determination of absolutely valid individuality, while the end of ethics was absolutely valid goodness and the end of aesthetics was absolutely valid beauty. The end of the discipline guided the formation of norms and concepts. But the ideal could never be fully actualized, for that would conflate validity with being. Ultimately, what the naturalist and the historicist lacked was "faith" in absolute values. In the ineliminable role of values—the primacy of practical reason—lay the core of Baden idealism.

On the Marburg view, psychologism's problem was not that it failed to acknowledge the irreducible duality of validity and being, but that it sought to overcome that duality in the wrong way. Philosophy was supposed to orient itself by *the factum of science*—or, better yet, *the fieri or being-made of science*—which consisted in the concept-formation whereby science's proper objects were constituted as *exceptionless laws*.[44] Thanks to mathematical and scientific advances, Kant's transcendental logic could be purified of everything providing a temptation to psychologize—including a priori intuition. Philosophy depended on the special sciences, yet philosophy alone could locate the proper *factum* or *fieri* of science for a specific domain. For example, the proper *factum* of the sciences of the human was to be found, not in morality or psychology, but in jurisprudence, where alone the concept of law-governed agency could be adequately formed.[45] Here idealism meant the methodological constitution of scientifically intelligible reality.

Marburg and Baden Neo-Kantians alike began by distancing themselves from both psycho-physiological Neo-Kantianism and Post-Kantian Idealism. Each criticized the other for repeating the errors of the Idealists. Thus Marburg Neo-Kantians found the Baden emphasis on normativity and value, as well as the ongoing role ascribed to the intrinsically irrational realm of non- or pre-scientific being, to be overly subjective, while Baden Neo-Kantians thought that the Marburg dismissal of the realm of being as irrelevant to science, along with the development of a logic whose principle of origin stated that, "Only thought itself can produce what can count as being",[46] was a revival of Hegelian panlogism that threatened to collapse into nihilism. But this reinforced the suggestion that affinities with Post-Kantian Idealism ran deep, since Marburg repeated Hegel's criticism of Fichte, while Baden reiterated Jacobi's and Schelling's indictment of Hegel. In the 1900s, these affinities were explored in a more open-minded fashion, as Cassirer, from the Marburg school, developed an account of both symbolic logic and Einstein's relativity theory, incorporating them within a philosophy of symbolic forms whose ambitions rivalled Hegel's *Phenomenology of Spirit*,[47] and Windelband, from the Baden school, announced that the "return to Kant" had become a "renewal of Hegelianism" in the quest for "a comprehensive philosophy of culture" that avoided, not only psychologism, but also historicism.[48] On the other hand, by rejecting Kant's notion of an a priori form of sensible intuition, both schools seemed to have exacerbated the duality that had prompted Kant to develop a transcendental deduction in the first place. If reality in itself was value-free, as the Baden school thought, or if everyday sense perception lacked all scientific relevance, as the Marburg school taught, then how was rational responsiveness to reality so much as possible?

[44] Natorp (1887). [45] Cohen (1904). [46] Cohen (1902), 81.
[47] Hegel (1807); Cassirer (1923–9), (1995). [48] Windelband (1910).

9. Conclusion

Naturalism and historicism still exert powerful forces over the philosophical imagination, while the methods of natural and human science continue to change and develop. So it should be no surprise that the project sketched above, which attempts to save reason from naturalist scepticism and historicist nihilism while accounting for the success of the sciences, remains vital. Much has been learned from Post-Kantian Idealist and Neo-Kantian expansions of Kantian methods. The question remains, however, whether these expansions might yet succeed in pre-empting the problem to which Kantian transcendental deductions are addressed—the problem of the duality of our sense-based and reason-based ways of understanding the world—or whether what is called for is, once again, a return to Kant.

References

Beaney, Michael, ed. (2007), *The Analytic Turn: Analysis in Early Analytic Philosophy and Phenomenology*. London: Routledge.

Cassam, Quassim (2007), *The Possibility of Knowledge*. Oxford: Clarendon Press.

Cassirer, Ernst (1923–9), *Philosophie der symbolischen Formen*. Three volumes. Berlin: Bruno Cassirer. Trans. Ralph Manheim (1955–98), *Philosophy of Symbolic Forms*. New Haven, CT: Yale University Press.

Cassirer, Ernst (1995), *Zur Metaphysik der symbolischen Formen*. Berlin: Felix Meiner. Trans. John Michael Krois and Donald P. Verene (1998), *Philosophy of Symbolic Forms. Volume 4. The Metaphysics of Symbolic Forms*. New Haven, CT: Yale University Press.

Cavell, Stanley (1972), *The Senses of Walden*. New York, NY: Viking Press.

Cavell, Stanley (1979), *The Claim of Reason*. Oxford: Oxford University Press.

Cohen, Hermann (1902), *Logik der reinen Erkenntniss*. Berlin: Bruno Cassirer.

Cohen, Hermann (1904), *Ethik des reinen Willens*. Berlin: Bruno Cassirer.

Conant, James (2004), "Varieties of Scepticism" in *Wittgenstein and Scepticism*, ed. Denis McManus, 97–136, London: Routledge.

Deleuze, Gilles (1984), *Kant's Critical Philosophy*. Trans. Hugh Tomlinson and Barbara Habberjam. London: Athlone Press.

Fichte, Johann Gottlieb (1794), *Grundlage der gesamten Wissenschaftslehre*. Trans. Peter Heath and John Lachs (1982), *The Science of Knowledge*. Cambridge: Cambridge University Press.

Fichte, Johann Gottlieb (1978), *Wissenschaftslehre Nova Methodo. Gesamtausgabe 4:2. Kollegnachschriften 1796-8*. Eds. Hans Glitwitzky and Reinhard Lauth. Bad Canstatt: Frommann-Holzboog.

Förster, Eckart (2012), *The Twenty Five Years of Philosophy*. Cambridge, MA: Harvard University Press.

Franks, Paul (2005), *All or Nothing: Systematicity, Transcendental Arguments, and Skepticism in German Idealism*. Cambridge, MA: Harvard University Press.

Franks, Paul (2007), "From Quine to Hegel: Naturalism, Anti-Realism and Maimon's Question *Quid Facti*" in *German Idealism: Contemporary Perspectives*, ed. Espen Hammer, 50–69, London: Routledge.

Friedman, Michael (1992), *Kant and the Exact Sciences*. Cambridge, MA: Harvard University Press.

Fries, Jakob Friedrich (1807), *Neue Kritik der Vernunft*. Heidelberg: Mohr und Zimmer. Revised edition (1828–31), *Neue oder anthropologische Kritik der Vernunft*. Heidelberg: Christian Friedrich Winter.

Hegel, Georg Wilhelm Friedrich (1807), *Phänomenologie des Geistes*. Bamberg and Würzburg: Joseph Anton Goebhardt. Trans. A. V. Miller (1979), *Phenomonology of Spirit*. Oxford: Oxford University Press.

Hegel, Georg Wilhelm Friedrich (1812–16), *Wissenschaft der Logik*. Nürnberg: Johann Leonhard Schrag. Trans. George di Giovanni (2010), *The Science of Logic*. Cambridge: Cambridge University Press.

Helmholtz, Hermann von (1850–67), *Handbuch der physiologischen Optik*. Leipzig: Voss. Trans. James P. Southall (1924–5), *Helmholtz's Treatise on Physiological Optics*. Menasha, WI: Optical Society of America.

Humboldt, Wilhelm von (1822), *Über die Aufgabe des Geschichtsschreibers, Abhandlungen der Historisch-Philosophischen Klasse der Königlichen Preussischen Akademie der Wissenschaften aus den Jahren 1820-21*. Trans. (1967), "On the Historian's Task". *History and Theory*, vol. 6, no. 1, 57–71.

Jacobi, Friedrich Heinrich (1785/1789), *Über die Lehre des Spinozas in Briefen an den Herrn Moses Mendelssohn*. Hamburg: Gottlob Löwe. Trans. George di Giovanni (1994), *Jacobi: Main Philosophical Writings*. Montreal, QC and Kingston, ON: McGill-Queen's University Press.

Jacobi, Friedrich Heinrich (1799), *An Fichte*. Hamburg: Friedrich Perthes.

Kant, Immanuel (1781/87), *Kritik der reinen Verunft*. Riga: Johann Friedrich Hartknoch. Trans. Paul Guyer and Allen W. Wood (1998), *Critique of Pure Reason*. Cambridge: Cambridge University Press. Cited by A/B (1781/87) pagination.

Kant, Immanuel (1788), *Kritik der praktischen Vernunft*. Riga: Johann Friedrich Hartknoch. Trans. Mary J. Gregor (1996), *Critique of Practical Reason*. Cambridge: Cambridge University Press. Cited by volume and page of *Kants gesammelte Schriften* (1900–). Berlin: Georg Reimer, later Walter de Gruyter.

Kant, Immanuel (1790), *Kritik der Urteilskraft*. Trans. Paul Guyer and Eric Matthews (2000), *Critique of the Power of Judgment*. Cited by volume and page of *Kants gesammelte Schriften* (1900-). Berlin: Georg Reimer, later Walter de Gruyter.

Körner, Stephan (1967), "The Impossibility of Transcendental Deductions". *The Monist*, vol. 51, no. 3, 317–31.

Lotze, Rudolf (1874), *Logik*. Leipzig: Hirzel. Trans. Bernard Bosanquet (1884), *Logic*. Oxford: Clarendon Press.

Maimon, Salomon (1790), *Versuch über den Transscendentalphilosophie*. Berlin: Christian Friedrich Voss. Trans. Nick Midgley, Henry Somers-Hall, Alistair Welchman, and Merten Reglitz. London: Continuum, 2010.

Müller, Johannes (1826), *Zur vergleichenden Physiologie des Gesichtssinnes des Menschen und der Tiere*. Leipzig: Cnobloch.

Mueller, Gustav E. (1958), "The Hegel Legend of 'Thesis-Antithesis-Synthesis'", *Journal of the History of Ideas*, vol. 19, no. 3, 411–14.

Natorp, Paul (1887), "Über objective und subjektive Begründung der Erkenntniss", *Philosophische Monatshefte*, vol. 23, 257–86. Trans. Lori Phillips and David Kolb (1981),

"On Objective and Subjective Grounding of Knowledge". *British Journal of the Society of Phenomenology*, vol. 12, no. 3, 245–66.

Schelling, Friedrich Wilhelm Joseph (1856–61), *Sämmtliche Werke*, division II, vol. 3, 74–94. Trans. Bruce Matthews (2007), *The Grounding of Positive Philosophy*, 141–54. Albany, NY: SUNY Press.

Stern, Robert (1999), ed., *Transcendental Arguments: Problems and Prospects*. Oxford: Oxford University Press.

Stern, Robert (2000), *Transcendental Arguments and Scepticism: Answering the Question of Justification*. Oxford: Clarendon Press.

Strawson, P. F. (1966), *Bounds of Sense: An Essay on Kant's Critique of Pure Reason*. London: Methuen.

Strawson, P. F. (1992), *Analysis and Metaphysics: An Introduction to Philosophy*. Oxford: Oxford University Press.

Stroud, Barry (1967), "Transcendental Arguments". *Journal of Philosophy*, vol. 65, no. 9, 241–56.

Stroud, Barry (1994), "Kantian Argument, Conceptual Capacities, and Invulnerability", in *Kant and Contemporary Epistemology*, ed. Piero Parrini, 231–51. Dordrecht: Kluwer.

Windelband, Wilhelm (1884), *Präludien*. Freiburg: Paul Siebeck. Expanded editions (1907, 1911, 1915, 1924). Tübingen: J. C. B. Mohr.

Windelband, Wilhelm (1910), *Die Erneuerung des Hegelianismus*. Heidelberg: Carl Winter.

CHAPTER 5

LOGICAL EMPIRICISM

CHRISTOPHER PINCOCK

> Neatness and clarity are striven for, and dark distances and unfathomable depths rejected. In science there are no "depths"; there is surface everywhere: all experience forms a complex network, which cannot always be surveyed and can often be grasped only in parts. Everything is accessible to man; and man is the measure of all things.
>
> Neurath, Hahn, and Carnap, "The Scientific Conception of the World: The Vienna Circle" (1929)

1. Introduction

Logical empiricism was a movement in philosophy that sought to transform many aspects of philosophy. This transformation aimed to change not only the philosophical claims that philosophers debated, but the very methodology of philosophy itself. By "methodology" I mean the choice of problems for philosophers to pursue and the acceptable methods for solving these problems. A methodological claim is of course a philosophical claim, but it is a claim about philosophy itself in a special way that distinguishes it from more ordinary philosophical claims about knowledge or existence. One of the more remarkable features of logical empiricism is its explicit methodological debates. Many logical empiricists showed a keen awareness of the need for a new conception of philosophical problems and methods, and this prompted several sharp disputes among the logical empiricists. A reconsideration of these discussions may be productive today precisely when a new round of methodological debates is underway among philosophers. While it is too much to hope that the logical empiricists' philosophical methodologies will provide a viable solution to the methodological problems of our own time, their debates remain instructive for philosophers today. Logical empiricism shows the value of a healthy skepticism towards the depth and profundity of philosophy itself, and it seems that this skepticism is largely absent in contemporary philosophy.

I will organize my exposition around three problems that preoccupied logical empiricism. Each of these problems was more or less ill-defined in its original form. The refinement of the problem itself and associated claims about the methods needed to resolve the problem prompted three rounds of methodological debate. The primary problem is commonly referred to as the coordination problem. Given that the sciences successfully deploy increasingly abstract tools, how should our scientific claims be coordinated with what we can experience and test? Debates about the coordination problem resulted in two additional problems. First, there is a problem about logic. Logic may provide the resources to account for the coordination of abstract structures with concrete reality, but what is logic and how should this sort of logical coordination be effected? Second, there is an alternative range of issues tied to the role of context in science. Scientific claims are the products of human activity carried out in a broader social and political context. What role should this context play in coordinating abstract structures with concrete reality?

Schlick and Reichenbach offered similar proposals for how to resolve the coordination problem in the earliest stages of logical empiricism. But they quickly parted ways based on their differing views of the significance of logic for coordination. Under the influence of Wittgenstein's *Tractatus*, Schlick came to believe that logic revealed the essential features of all representations and that logical tools were both necessary and sufficient to resolve the basic coordination problem. Reichenbach, by contrast, appealed instead to substantial claims about causality and probability. Carnap in the *Logical Syntax of Language* developed yet another methodological option based on his logical pluralism and conception of scientific knowledge as tied to a formally articulated language. Finally, Neurath and Frank responded by appealing to features of the context in which science is done. This required a different conception of the coordination problem and its solution. While logic was assigned some role in coordination, greater emphasis was placed on the actions of individuals and their social context. A reply to this expansive conception of philosophy can be found in Reichenbach's celebrated distinction between the contexts of discovery and justification. After surveying these debates, I will conclude with some remarks about the methodological proposals offered by Hempel in his later work.[1]

2. Coordination

Nineteenth-century successes in physics and chemistry, combined with a new experimental psychology, prompted many to worry that it was no longer clear what the sciences were studying or how our scientific knowledge was to be justified. Several writers traced these worries to an unreasonable tendency to hold on to traditional common sense and

[1] The contemporary scholarship on logical empiricism is considerable. Important introductions are Stadler 2001, Richardson 2003, Rescher 2006, Richardson and Uebel 2007 and Friedman and Creath 2007. I am especially indebted to the work of Michael Friedman and Thomas Uebel. See in particular, Friedman 1999, Friedman 2000, and Uebel 2007. I am also grateful to Greg Frost-Arnold, Robert Kraut, and Gregory Lavers for the opportunity to review some of their unpublished work on logical empiricism. See in particular, Frost-Arnold 2013, Kraut forthcoming, and Lavers forthcoming. I would also like to thank the editors and an anonymous referee for helpful suggestions.

religious beliefs in the face of their lack of scientific standing. A forceful instance of this trend is Ernst Mach. Mach sought to recast the scientific claims of both psychology and physics in terms of neutral elements. At the same time Mach recognized the need to bring to bear some conceptual organization of these elements if our cognition was to function. He insisted only that these concepts prove their scientific value. The assumption was clearly that most traditional beliefs would not survive this sort of scrutiny and would be replaced by a new conception of the world: "In place of the traditional, instinctive ways of thought, a freer, fresher view, conforming to developed experience, and reaching out beyond the requirements of practical life, must be substituted throughout" (Mach 1897, p. 24).[2] It is important to see Mach's empiricism in this light. Mach's empiricism was not the philosophical claim that all knowledge is justified by experience. It was instead closer to what van Fraassen has termed the empirical stance: an attitude undertaken at a given time for a particular purpose (van Fraassen 2002). Mach's goal was to overcome the tensions between the sciences and our non-scientific beliefs.

Moritz Schlick arrived in Vienna in 1922 as the professor of philosophy of the inductive sciences, occupying a chair once held by Mach. Schlick embraced the idea that our beliefs must withstand critical scrutiny, especially in light of the development of Einstein's general theory of relativity. One theme of Schlick's *Space and Time in Contemporary Physics* is that our ordinary beliefs about space and time reflect inherited biases that must be swept away. At the same time, Schlick was highly critical of Mach's immanence philosophy: the view that all that exists is part of someone's experience. In *General Theory of Knowledge*, Schlick defended a form of scientific realism. The crucial means to this conclusion is Schlick's novel account of what scientific concepts are and how they might be connected with reality. He adapted Hilbert's account of the implicit definition of geometrical concepts by axioms to all mathematical and scientific concepts. On this approach, a scientific theory begins as a sequence of statements. The non-logical terms of these statements are uninterpreted, but the collection of statements is taken to define what those non-logical terms mean. This seemed to work especially well for terms for distances and times as well as fundamental physical quantities like mass and charge. To arrive at a genuine theory, the uninterpreted statements must be related or coordinated with empirical reality. Crucially, Schlick assumed that these coordinations must involve the stipulation that something observed is the referent of the implicitly defined terms. This comes through clearly in Schlick's emphasis on point coincidences: "The adjustment and reading of all measuring instruments of whatever variety … are always accomplished by observing the space-time coincidence of two or more points" (Schlick 1963, pp. 49–50). For example, the use of a clock to measure a period of time involves the coincidence of the hands of the clock with marks on the clock. So, it is possible to use observations of a clock to stipulate what some theoretical terms for times refer to.

Given this sparse observational basis, Schlick was forced to coordinate his abstract terms with very simple observational phenomena. The coordination problem was to be solved, then, by finding suitable observational phenomena for sufficiently many abstract scientific concepts. On this approach, the philosopher of science has two primary tasks: first, to clarify the different elements that must be addressed to solve the coordination problem;

[2] This passage is partially given in Nemeth 2007.

second, to specify how particular theories should be reconstructed in these terms and how those theories might fit together. A secondary task is to combat alternative approaches that invoked superfluous machinery. *General Theory of Knowledge* aims to accomplish both tasks by developing a conception of knowledge that goes beyond Machian positivism and yet takes seriously the conception of knowledge involved in this sort of coordination (Schlick 1985).

While Schlick was developing his account of the coordination problem, Hans Reichenbach tackled similar issues in Berlin. Also preoccupied with the special and general theories of relativity, Reichenbach distinguished axioms of connection and axioms of coordination. The former played the same role as Schlick's implicit definitions while the latter served to link uninterpreted non-logical terms to concrete reality. But Reichenbach sought to give more substance to his axioms of coordination. Early on, Reichenbach gave axioms of coordination a constitutive role that shared some features with Kant's synthetic a priori. These axioms were revisable, unlike Kant's synthetic a priori judgments, but at the same time they were "constitutive of the concept of the object" (given by Friedman 2007, p. 98). As with many Kantian claims, this notion of constitution can be given a more idealistic or more realistic interpretation. On the idealistic interpretation, constitution requires that the object be mind-dependent and somehow created by the scientific community. As axioms of coordination are revisable, this entails that what exists is also changing over time. The more plausible realistic interpretation is that it is only our concepts that are changing, but that the concepts at a given time will pick out genuine objects if our definitions succeed in matching the right features of those objects. On this approach we wind up with a genuine scientific theory if we supplement our axioms of connection with the right sort of coordinating axioms. The job of the philosopher of science is to figure out what these axioms should be and to reconstruct our scientific theories in these terms. This project is pursued in Reichenbach's *The Philosophy of Space and Time*. While insisting that axioms of coordination are merely definitions, Reichenbach also grants the need to invoke approximations and limiting procedures when defining crucial notions like the concept of a rigid body (Reichenbach 1958, p. 22).

While Schlick's and Reichenbach's conception of the coordination problem are closely linked, even in the 1920s there were signs of a crucial disagreement. From Schlick's starting point, it is tempting to assume that coordination is an act of stipulation that lies at the basis of all of our scientific knowledge. It seems to follow that a particular act must involve only what is given independently of all scientific knowledge. This restricts the primary referents of the non-logical scientific terms to Mach's neutral elements or some similarly unconceptualized experience. Reichenbach is able to avoid this extreme conclusion by conceiving of the coordinating axioms as substantial claims about a realistically construed physical world. The claims are substantial because they could be false. For example, the definition of rigid bodies makes use of a distinction between internal and external forces, and this in turn assumes that sense can be made of a closed system. Unfortunately, "a closed system can never be strictly realized" (Reichenbach 1958, p. 23), and so some limiting procedure must be invoked that assures us that our purported rigid body would behave in a certain way in a series of increasingly isolated systems. The validity of this limiting procedure is a risky claim about the world. The challenge for Reichenbach is to explain how we can support these sorts of claims. Ultimately Reichenbach must posit certain highly general

regularities tied to causation and probability. These regularities lie at the heart of our scientific theories, but Reichenbach has difficulty accounting for our knowledge of them.

Rudolf Carnap's *Logical Structure of the World* (also known simply as the *Aufbau*) (Carnap 2003) is sometimes read as a proposed solution to Schlick's coordination problem. On this interpretation the *Aufbau* gets by with a single act of coordination between the relation symbol "Rs" and the relation of recollected similarity on the total momentary experiences of an individual. The definitions of scientific concepts that Carnap offers as part of his "autopsychological" constitution system are then analogous to Schlick's implicit definitions. On an alternative interpretation, which I prefer, Carnap's constitution system is offered instead as a rational reconstruction of our scientific knowledge.[3] Taking our scientific knowledge for granted, Carnap reorders it to reflect relationships of epistemic priority. The acts of coordination emphasized by Schlick and Reichenbach thus find a place within the constitution system itself (Friedman 2007, p. 208). There is no account of how our words hook on to reality as the notion of a system-independent reality is dismissed as metaphysical. This interpretation of Carnap's *Aufbau* suggests a somewhat different methodology for philosophers. Now the aim is to build a unified conceptual system within which all of our concepts can find a place. Some of these efforts will resemble the implicit definitions or coordinative definitions of Schlick and Reichenbach. But there is no attempt to transcend the system itself, and so coordinative definitions in their objectionable ostensive or realistic forms are avoided. Unfortunately, in the *Aufbau* itself Carnap devoted little space to indicating the nature of logic or how logical structure might be isolated. Without these clarifications, it remained unclear how merely arranging our concepts in a logical system could address the worries that motivated the original coordination problem. Investigations of the nature and philosophical significance of logic thus formed the next phase of logical empiricism's methodological development.

3. Logic

In its early stages, most logical empiricists were content to rely on the new logic of Frege, Russell, and Hilbert without placing the nature of logic on the philosophical agenda. The central role of logic in discussions of coordination put pressure on this methodological position. The pressure is clear in the unsatisfactory discussions of logic in the works by Schlick, Reichenbach, and Carnap just reviewed.[4] The situation changed considerably when logical empiricists turned to a careful study of Wittgenstein's *Tractatus* (Wittgenstein 1974). Schlick and Carnap developed quite different conceptions of the role of logic in philosophy in this period. These differences showed themselves clearly in a debate about the correct way to present the most basic elements of our knowledge. This is the notorious "protocol sentence" debate of the 1930s. We will return to Reichenbach at the end of section 4.

Schlick coopted a number of Tractarian doctrines and embraced a form of verificationism that displaced the limited scientific realism of *General Theory of Knowledge.* A concise

[3] See Pincock 2009 for a survey of different interpretations.

[4] Schlick 1985, part II, Carnap 2003, section 107, and Reichenbach 1958, section 7.

formulation of this new position is to be found in "The Turning Point in Philosophy" (Schlick 1930), the paper that opens the newly launched journal *Erkenntnis*. Announcing the end of "the fruitless conflict of systems" (Schlick 1930, p. 54), Schlick extolled "methods which make every such conflict in principle unnecessary" (Schlick 1930, p. 54). According to Schlick, Wittgenstein has determined the nature of logic by focusing on what is essential to all representations. He agrees with Wittgenstein that this form cannot be represented. So, the task that is left to philosophy is merely one of clarification. This activity is ultimately the assignment of meanings to linguistic items. These meanings cannot be assigned by using statements, but the process "always comes to an end in actual pointings, in exhibiting what is meant, thus in real acts" (Schlick 1930, p. 57). It is clear that Schlick is here building on his earlier focus on coordinative definitions. But the new Wittgensteinian conception of logic closes off any further role for philosophy. Schlick is now also clear that these acts must involve only what is given to us in immediate experience. This makes his earlier realism unattainable, for our words can only mean what we point to, and Schlick assumes we can only point to our immediate experiences.

Schlick can be seen to illustrate one extreme in the methodological development of logical empiricism. It seems obvious that one role for philosophy is to clarify claims, but Schlick's view that this is the only possible philosophical activity clashes with the wide variety of activities that have been historically ascribed to philosophers. Furthermore, Schlick's insistence that the only material available in "actual pointings" is immediate experience is unmotivated. By the time of "The Foundation of Knowledge" (Schlick 1934), the untenability of Schlick's account of philosophy was becoming obvious. There Schlick places a special sort of statement that he calls a "confirmation" at the foundation of our knowledge.[5] Such statements are foundational because "I grasp their meaning at the same time as I grasp their truth" (Schlick 1934, p. 225). To play this role, they must pertain only to purely demonstrative matters. All ordinary statements are thus merely hypotheses that are confirmed to some extent by the confirmations that they generate. Scientific knowledge, then, rests on a slender basis of private experiences of satisfaction. Philosophy aims to isolate this slender basis and force scientific theories into this rigid Wittgensteinian framework.

Carnap proceeded in a completely different direction and developed what was to become the most lasting methodological innovation of the logical empiricists. This is the principle of tolerance and the corresponding conception of philosophy and its methods. Carnap began this phase of his development by embracing the view of the *Tractatus* that there was only one logic that reflected the essential features of any system of representation. This position is found in the influential "The Elimination of Metaphysics Through Logical Analysis of Language" (Carnap 1932). There Carnap criticized Heidegger for violating the rules of logical syntax. This criticism is only coherent if there actually is a single system of rules for logical syntax that we can hold up as a standard against which to measure the statements of others. But soon after this paper Carnap accepted the more tolerant position: "In logic, there are no morals. Everyone is at liberty to build up his own logic, i.e. his own form of language, as he wishes. All that is required of him is that, if he wishes to discuss it, he must state his methods clearly, and give syntactical rules instead of philosophical arguments" (Carnap 2002, p. 52). If one accepts the principle of

[5] The German word "Konstatierung" is also translated as "affirmation."

tolerance, then it is still possible to criticize Heidegger for failing to "state his methods clearly" as the proposed logical syntax of Heidegger's language is far from clear. But this sort of criticism is different from the dogmatic insistence that one's opponent is engaged in meaningless metaphysics.

Carnap's path to tolerance is a complex one (Awodey and Carus 2007). Even in *Aufbau* Carnap had allowed several systems of definitions of our scientific concepts ("constitution systems"). He seems to have originally aimed for a single logical framework that would allow the comparison and investigation of these systems and the associated axiom systems of mathematics. However, Gödel's incompleteness theorems eventually convinced Carnap that this sort of all-embracing metalanguage was not achievable. The alternative was to allow that any language could be presented as a candidate for the language of science. To propose a language one must specify the vocabulary along with the formation rules for sentences and the transformation rules for inferences. In *Logical Syntax of Language* Carnap offers an elaborate procedure that will allow one to use these rules to isolate the logical expressions along with those sentences that are L-valid or analytic for a given language (Carnap 2002, sections 50–2). Carnap clearly hoped that the logical and mathematical statements of his languages would wind up as analytic. Strictly speaking, though, there is no prior language-independent way to pick out the mathematical statements of a language to check whether this aim is realized. All that is available is the language-relative distinction that results from Carnap's definitions.

It is difficult to appreciate how Carnap sought to use these proposed languages to address anything like our original coordination problem. The challenge is to link a descriptive expression of some proposed language to an experience or object in the physical world. Without such a link, it seems that Carnap's languages are purely formal entities that are poor candidates to be vehicles for genuine scientific knowledge. Carnap encourages this interpretation by insisting that all of philosophy is merely the investigation of the logical syntax of these languages. In restricting the focus of philosophy to logical syntax, Carnap aims to exclude questions about reference, meaning, and truth. To the extent that such semantic questions are legitimate, it must be possible to translate the "material mode" semantic question into an equivalent "formal mode" syntactic question. For example, a statement in the material mode about reference could be equivalent to a formal mode statement about the occurrence of a word or its synonyms, where synonymy is treated in terms of the syntactical rules of the language in question.

Carnap grants that one could present a syntactic definition of a kind of "protocol" sentence that would serve a special role in the testing and confirmation of scientific hypotheses. However, his position in *Logical Syntax* appears to be that the philosopher is not tasked with determining which protocol sentences one should assert: "the statement of the protocol-sentences is the affair of the physicist who is observing and making protocols" (Carnap 2002, p. 317). This is quite different from Schlick's attitude. Schlick sought a special connection to experiences that would privilege the protocol sentences and explain their special epistemic role. Carnap instead delegates the choice of protocol sentences to the scientist. It is only within a given language that we can make acceptable inferences, and so it appears that Carnap has nothing to say about which protocol sentences are correct. Carnap is more forthcoming in "Testability and Meaning" (Carnap 1936). There he supplements his "Logical Analysis of Confirmation and Testing" with an "Empirical Analysis of Confirmation and Testing". It looks like he has decided that empirical matters are properly

philosophical as he devotes 17 pages to discussing them in one of his philosophy papers. But when we turn to the contents of that section, we again find Carnap deferring to scientific findings. Two basic terms "observable" and "realizable" are assumed and used to define other terms like "confirmable". But the basic terms are defined "within psychology, and more precisely, within the behavioristic theory of language" (Carnap 1936, p. 454). This marks a clear boundary between the formal, logical studies of the philosopher and the related, though distinct, empirical work of psychologists and linguists.

Carnap's mature conception of philosophy retains tolerance while allowing semantic concepts to play a part in genuinely philosophical projects. The main impetus for this shift was Tarski's theory of truth. Tarski showed how the concept of truth for a given language L1 could be defined in a metalanguage L2 if L2 was granted certain expressive resources. This convinced Carnap that semantic notions such as reference, meaning, and truth were no more problematic than the syntactic notions deployed in *Logical Syntax*. After this shift Carnap was forced to repeatedly make clear that his notions of meaning and truth had none of the traditional metaphysical overtones that critics worried about. Notoriously, Ryle complained in his review of *Meaning and Necessity* that Carnap had assumed a discredited " 'Fido'-Fido" conception of language (Ryle 1949).[6] Carnap responded in "Empiricism, Semantics and Ontology" (Carnap 1950a) that his semantic notions are internal to a given metalanguage, and are not intended to take a stand on the traditional metaphysical question of how words hook on to objects in some mind-independent reality. Existence claims thus come in two flavors. If one asserts "There are Fs," then this could be meant as a claim internal to a given language. In the special case where the existential claim follows simply from the rules of the language in question, then the claim is analytic and the existence of these entities is guaranteed for anyone who adopts that language. This was Carnap's attitude towards numbers and semantic entities like propositions. In other cases where "There are Fs" is a well-formed sentence of the language, but neither it nor its negation follows from the rules, then we have an ordinary empirical claim whose status must be resolved by ordinary scientific methods. But if the claim "There are Fs" is meant as an external claim, then Carnap denies that such claims make any sense: "An alleged statement of the reality of the system of entities is a pseudo-statement without cognitive content" (Carnap 1950a, p. 214). All that can be accomplished by such an external claim is that the speaker makes a proposal for what sort of language to adopt.[7] And for Carnap the proper language to adopt is determined only by one's practical purposes such as building a rocket or feeding the inhabitants of a city. So, Carnap was not only opposed to the " 'Fido'-Fido" theory of language that Ryle saw in *Meaning and Necessity*, but was of the opinion that any such theory was theoretically meaningless.

Even after the turn to semantics, then, the methods available to the philosopher qua philosopher are quite restricted. A philosopher can propose and investigate formal languages. These languages or linguistic frameworks are individuated by their syntactic and semantic rules. It is possible for the philosopher to distinguish certain terms as observable and the rest as theoretical, and to investigate the logical relationships between sentences formed using these terms. In large measure, these languages are developed with an eye on the needs

[6] On this view all meaningful terms are thought to be names. See also Eklund 2010.

[7] See Price 2010 and Kraut forthcoming for two elaborations of this view.

of science. Ultimately the goal is to find a formal language that could serve as the language of a unified science that would incorporate all the various branches of science into one precise system. Unsurprisingly, Carnap never came close to achieving such a language himself. He spent much of the last phase of his career considering the problem of probability and confirmation. At the same time, he tried to refine the logical tools that the philosopher could use to develop new languages. Carnap insisted that the philosopher cooperate with the scientist when executing philosophical tasks. This engagement takes two forms. First, the philosopher should find out from the scientists what sort of scientific theories are being developed, and use this information to focus his or her philosophical efforts on languages of the appropriate form. Second, the philosopher should cooperate with the linguist to see how the formal languages found in philosophy can be mapped on to the linguistic behavior of actual speakers. This sort of "pragmatics" of language is not part of philosophy, but without having an eye on what speakers actually do, and how their behavior could be reformed to better achieve practical goals, the formal activity of the philosopher risks being pointless (Ricketts 2003).[8]

Carnap's conception of philosophy and its methods achieved a kind of internal coherence that has continued to attract those skeptical of traditional metaphysics or impatient with never-ending philosophical disputes. For a dedicated Carnapian, there are no genuine philosophical disagreements, but only unclarities tied to the failure to articulate one's linguistic framework. The optimistic attitude is that if we were only to take the time to make these commitments explicit, then endless philosophical debates would be ended. For the non-Carnapian, however, the conception of philosophy offered by the Carnapian can seem unduly restrictive. For we typically intend to make factual claims, and the Carnapian cannot countenance such claims in the domain of philosophy. And most philosophers aim to deploy methods quite different from merely proposing and analyzing linguistic frameworks.

Two large challenges face the Carnapian conception of philosophy. They are what I will call the neutrality problem and the naturalism problem.[9] Recall that the principle of tolerance insisted that we are free to choose the logic of our language. This suggests that Carnapians should occupy a neutral vantage point from which they can assess the available options for the logic of the proposed language of science. At the same time, it is clear that a metalanguage must be chosen for this sort of comparison to even get started. The neutrality problem is that the choice of metalanguage will bias our choice for the language of science and thus deprive the Carnapian of their neutrality (Friedman 1999, ch. 9, esp. sections VI and VII). For example, if we choose a metalanguage with classical logic, then any worries about classical logic for the object language will be easy to address. What this neutrality problem highlights is that even the most dedicated Carnapian must recognize a role for decisions that are not subject to any further rational scrutiny. It might seem like

[8] Uebel 2013 ascribes a "bipartite metatheory" of science to Carnap that includes these pragmatic investigations. However, Uebel is in general agreement with our interpretation of Carnap as he is clear that pragmatic investigations, though important, lie outside of the scope of philosophy: "Carnap was more concerned to distinguish his logic of science from these empirical studies than to pursue them" (Uebel 2013, p. 526).

[9] A third problem tied to the limits of formalization will be discussed in section 5.

everything is open to examination and stands to be tested by its efficiency as a means to our chosen ends. But this sort of neutrality is impossible to achieve.

The naturalism problem is known to most contemporary philosophers in the form raised by Quine. We have seen that Carnap's entire conception of philosophy depends on the distinction between statements that are licensed by the choice of the language and statements that are licensed by other reasons. The naturalist challenge rejects this distinction. For the naturalist (as I will use this term), at any given time our choice of which statements to accept is determined both by the language we have adopted and by our empirical evidence. It is not possible or desirable to try to separate our choices into the two factors that Carnap identifies. Quine raised the naturalism problem from the narrow perspective that insisted that only the physical sciences had standing in our philosophical deliberations (Creath 2007). However, there is no need for the naturalist to accept this restriction. One could instead begin with the broad assumption that both the physical and social sciences were equally deserving of our respect. The resulting conception of philosophy is to be found in Otto Neurath. We can take Neurath's naturalism to involve a different solution to our original correspondence problem.

4. Context

For Neurath, the solution to the correspondence problem lies not in the nature of logic, but in the activities of human agents engaged in the task of navigating a complex world. Scientists are already in contact with the world when they use language to communicate with one another and deploy measuring instruments to physically interact with the world. There is thus no need to coordinate the uninterpreted abstract statements of Schlick. Furthermore, Neurath was highly skeptical of the scientific significance of appeals to the private experiences of individuals. Science as observed proceeds by public statements and procedures whose reliability consists in the potential for scrutiny and checking by other scientists. Neurath's approach to philosophy emphasized all of these themes, although he often became sidetracked by the need to engage in polemic exchanges with other philosophers, including Schlick and Carnap.

Much of Neurath's philosophical methodology can be gleaned from the famous boat metaphor from "Protocol Sentences" (Neurath 1932):

> *There is no way of taking conclusively established pure protocol sentences as the starting point of the sciences.* No *tabula rasa* exists. We are like sailors who must rebuild their ship on the open sea, never able to dismantle it in dry-dock and reconstruct it there out of the best materials. Only the metaphysical elements can be allowed to vanish without trace. Vague linguistic conglomerations [Ballungen] always remain in one way or another as components of the ship.
>
> (Neurath 1932, p. 201)

A number of important points are combined in this one passage. To start, it is clear how hostile Neurath is to Schlick's approach to coordination. Schlick required an implicitly defined network of terms to be coordinated with the immediate experiences of individuals. This sets up Schlick's confirmations as "pure protocol sentences" that are entirely

responsible for the empirical content of our scientific statements. By contrast, Neurath insists that we are already using a language whose content is more or less acceptable as it stands. It remains imperative to reform this language, clarify its overall structure, and purify the statements we accept of needless accretions. But the reconstruction that Schlick offers is unable to assist in this process.

In his positive proposal for protocol sentences Neurath presented a stark alternative to Carnap's program as well. In the 1932 paper Neurath offered the following as a potential protocol: "Otto's protocol at 3:17 o'clock: [At 3:16 o'clock Otto said to himself: (at 3:15 o'clock there was a table in the room perceived by Otto)]" (Neurath 1932, p. 202). A charitable reconstruction of what Neurath intends here has been offered by Uebel (Uebel 2007, ch. 11). The basic idea of this interpretation is that Neurath is attempting to clarify the process of scientists offering and evaluating protocols or reports on their experimental observations.[10] A scientist might actually say something like "At 3:15 o'clock there was a table in the room." But Neurath notes that the scientific community does not accept this report without first carefully considering it. On Uebel's interpretation, the reformed protocol that Neurath offers is meant to tease out the various tests that the scientific community may impose before they will accept the embedded report. These tests are not merely formal, but also involve considerations tied to the social and political context. By making this critical process more explicit in the form of the proposed protocol statements themselves, Neurath hopes to aid the scientific process of testing and hypothesis refinement. At the same time, the proposal is meant to make clear that no appeal to private experiences or other philosophical refinements is needed to make sense of scientific testing and scientific knowledge more generally. Some philosophers might try to screen out the contextual factors that Neurath emphasizes, but Neurath shows no interest in removing these sorts of considerations. His hope seems to have been that the best kinds of tests that we can hope for are just those scientific procedures that we already have, once they are made clear and open to community-wide participation.

The outermost layer reflects the need to check if Otto is really offering the statement as a scientific report. If Otto is not a scientist, his protocol might be rejected even at this stage. The next layer corresponding to "Otto said to himself" focuses on the sincerity of Otto's protocol. He might, for example, be lying, and so it would not be right to say that he said the rest of the protocol to himself. Finally, there is the question of what was actually perceived by Otto. Here it might be that Otto is hallucinating or making some other sort of error in reporting his perception. If all these tests are met, then it is appropriate to add the embedded report to the body of statements accepted by the scientific community. Even here Neurath allows that there might be good reason to reject the report. This might happen if accepting the report would undermine a valuable store of accepted statements.

It should be clear how important the social context of the scientific process is to making sense of this procedure. At each stage the choice about how to proceed is not determined by any logical rules or formal procedure. There is ultimately only the judgments of individual scientists and their collective practices. As Neurath once cryptically put it, "a logically tenable multiplicity is reduced by life" (Neurath 1935, p. 117), where "life" includes

[10] So, Neurath is like a sailor attempting to rebuild his ship while at sea. He is a scientist who is trying to improve the practice of science by making these critical suggestions.

our historical and social situation. This has left the impression that Neurath mandated some kind of coherence theory of truth or worse, as with Russell's unfortunate remark that according to Neurath "empirical truth can be determined by the police" (Russell 1940, p. 140). To start, just as Carnap did not offer any traditional account of how words refer to things, Neurath did not offer any traditional theory of truth. His considered view is that the notion of truth was just one of those metaphysical concepts that the ship of science should discard (Mancosu 2008). Instead, Neurath was offering an account of statement acceptance, and in these terms he insisted that no algorithm for statement acceptance was available. This is because the factors that are relevant to accepting or rejecting a statement are not exhausted by elements that can be captured in the formally defined linguistic frameworks familiar from Carnap. As we saw earlier, Carnap attempted to consign these contextual matters to the "pragmatics" of language use and the engineering decision about which logical means are best to achieve our practical goals. Carnap separated out these contextual factors from the proper task of the philosopher. Neurath can be seen, then, as emphatically reintroducing these factors and arguing that they are indeed central to any appreciation of how science works. To the extent that this context includes history as well as social and political values, Neurath's proposal is quite radical.[11]

Neurath proposed a philosophical methodology, then, where the formal, logical studies of Carnap would live alongside a "behavioristics" of science that would describe and reform the behavior of the scientific community. Neurath clearly thought that the latter task was the primary one for it was the only one that would make contact with scientific knowledge as it actually is. This is one way to understand Neurath's work on the unity of science project. For many years Neurath sought to create a kind of encyclopedia that would bring together in a series of volumes the scientific knowledge of his day.[12] It would obviously have made little sense to have these volumes written in the logical languages that Carnap was developing. They were written in a natural language so that the findings of the special sciences could be absorbed and used by other scientists and even the general public. The political dimension of this project is hard to ignore. For Neurath, the ultimate goal was a widely accessible summary of natural science and social science that could be used to better educate individuals and allow enlightened policies tied to central planning and social progress.

Neurath's main ally in this project was Philipp Frank. In "The Place of Logic and Metaphysics in the Advancement of Modern Science" (Frank 1948) Frank complements Neurath's approach by arguing that metaphysics results from a failure to understand how the problem of coordination is to be solved (Uebel 2011). As with other logical empiricists, Frank insists that contemporary scientific theories deploy concepts that have little contact with common sense or ordinary experience. Problems arise when philosophers and scientists fail to recognize this break, and try to force scientific hypotheses into a predetermined interpretive framework:

[11] Quine construed the relevant context much more narrowly, and this marks a decisive difference between Quine's and Neurath's philosophical methodologies.

[12] See Cartwright, Cat, Fleck, and Uebel 1996 and Reisch 2007 for additional discussion and references.

> What we call in a vague way "common sense" is actually an older system of science which was dropped because new discoveries demanded a new conceptual scheme, a new language of science. Therefore the attempt to interpret scientific principles by "common sense" means actually an attempt to formulate our actual science by the conceptual scheme which was adequate to an older stage of science which is now abandoned.
>
> (Frank 1948, p. 285)

This attempt is bound to fail, and what we are left with are vague analogies between common-sense terms and the terms of the new science. This not only complicates any understanding of what the scientific hypotheses are actually aiming at, but blocks the necessary testing and revisions that constitute scientific progress.

In this diagnosis of the nature of metaphysics, with all of its historical and psychological aspects, Frank is exemplifying the broad scope of philosophy as he envisioned it. The methodological program that Frank sought to execute is described in papers like "The Institute for the Unity of Science: Its Background and Purpose" (Frank 1947). The role of philosophers goes far beyond merely articulating formal linguistic frameworks and seeing how these frameworks fit or fail to fit with scientific practice. The philosopher must also engage with the history and political context of science as it is actually done. This allows a more effective intervention in the practice of science so that the logical insights of the more restricted formal studies can be actually used to make science work better. The Institute that Frank describes will instill a "new logic of science" (Frank 1947, p. 163) that will supersede the outdated old logic of science that sought to interpret all science in Newtonian terms. This traditional approach answered the special sciences' search for unity by an artificial or superficial interpretation that failed to do justice to what each special science had accomplished. The first part of this project is a "logico-empirical or semantical analysis" (Frank 1947, p. 165) that will render each of the special sciences in an appropriately unified language. Here Frank cites Neurath's unity of science volumes as a successful step in this direction. However, a second part is essential: "In order to understand a particular choice of symbols logico-empirical analysis has to be complemented by an investigation of the psychological and sociological causes which have led to this choice" (Frank 1947, p. 166). This "socio-psychological analysis" is critical because only it will allow a true understanding of why science has developed in the way that it has.

Few would recognize this sort of contextually situated and politically motivated scholarship as properly philosophical. A canonical response to the programs of Neurath and Frank is offered by Reichenbach in the 1930s. As Reichenbach's views developed, he continued to defend a form of scientific realism that he argued was the natural development of logical empiricist ideals (Reichenbach 1936). The key move that Reichenbach argued for was the replacement of a verificationist theory of meaning with a theory of meaning based on probability assignments. While verificationism led to the sort of positivism defended by Schlick, the allowance for probability assignments as a form of meaning was a crucial first step for accepting meaningful claims about the future and unobservable entities. In the course of making this argument Reichenbach offered his famous distinction between the context of discovery and the context of justification. Epistemology aims to "construct justifiable sets of operations which can be intercalated between the starting-point and the issue of thought-processes, replacing the real intermediate links" (Reichenbach 1938, p. 5). Psychology aims merely to describe the actual processes of thinking without reference to their correctness. When we consider the context of justification of a given claim, though, we should consider only the evidence for this claim.

With the distinction between discovery and justification in place, Reichenbach argues that the philosopher's ultimate focus is rational reconstruction, and any interest in social and political factors in scientific decision-making is not properly philosophical.

It is important to emphasize that this distinction is merely the first step in Reichenbach's mature theory of philosophical methodology. He goes on to emphasize that epistemology has a significant descriptive component in addition to its critical component (Reichenbach 1938, p. 11). The description of scientific activity, when suitably reconstructed in the context of justification, reveals the need for many "volitional decisions" (Reichenbach 1938, p. 9) at the heart of science. The most important of these decisions fix the aims of science. Unfortunately, scientists and philosophers may misunderstand the relationship between the decisions that have been made, and so fall into a kind of incoherence. The "advisory task" (Reichenbach 1938, p. 13) of the epistemologist is thus to set forward coherent sets of decisions, principally in connection with the ends chosen for scientific activity.

This is how Reichenbach conceives of his theory of meaning in terms of probability assignments. It is a "given sociological phenomenon" (Reichenbach 1938, p. 146) that scientists make predictions about the future and act on these predictions by, for example, taking out life insurance policies. Reichenbach argues that a positivist theory of meaning in terms of conclusive verification cannot make sense of these actions. So, if we are to propose a language that might license these actions, then only a realistic language and the associated procedures for probability assignments are allowed. This is the set of coherent decisions that does the best job of making sense of our current use of words like "knowledge" and "prediction."

While Reichenbach's choice of language creates the space for justified beliefs about the future and unobservable entities, it is far from clear that he actually delivers the justifications. The main problem is that Reichenbach is not able to justify the probability assignments that go into his probability calculations. He is eager to defend a frequentist interpretation of probability. On this interpretation a probability claim is about the relative frequencies of certain outcomes over a repeated process. Reichenbach aims to use this very notion of probability in "projections" from the observed to the unobservable (Reichenbach 1938, pp. 124, 154–5). However, he is also aware that in many important cases the kinds of claims in question lack a sufficient track record to be placed in an appropriate reference class. Reichenbach's solution to this problem is to invoke what he calls a "blind posit" (Reichenbach 1938, p. 353). This is a bet or wager on a claim that is justified because it is the best means of achieving some end. These posits lie at the heart of Reichenbach's justification of induction and the associated probability assignments for scientific claims. The basic idea is that logical reasoning alone can assure us that a particular rule of induction is the best one available. It is not guaranteed to work, but will work if any such rule will work.

Reichenbach's project of vindicating scientific realism thus founders on two fronts. First, as the many discussions of his justification of induction have made clear, his choice of inductive procedure does not appear to have the privileged place he assigned it (Salmon 1991, Galavotti 2011). Second, even if the choice of the rule for induction could be defended, it remains unclear how this rule will translate into appropriate probability assignments for scientific theories. Until these problems can be overcome, one must judge Reichenbach's defense of scientific realism as incomplete.[13]

[13] Psillos 2011 provides a recent, critical reconstruction of Reichenbach's argument for scientific realism.

Despite its limitations, Reichenbach's conception of philosophy was quite influential, especially in the philosophy of science. In *How the Cold War Transformed Philosophy of Science* and elsewhere, Reisch has documented the extent to which Neurath's and Frank's more expansive conception of the philosophy of science was marginalized by more or less political considerations (Reisch 2005, 2007). From this perspective, the more technical programs pursued by Reichenbach and Carnap flourished in part because of their less threatening and more specialized character. One can grant Reisch's historical point, though, and still ask how we should evaluate these developments today. I have suggested that the philosophical methodology of Carnap is too restrictive and that Reichenbach was unable to justify his conception of our scientific knowledge using the resources he allowed himself. But this by itself does not vindicate the expansive programs of Neurath and Frank. Just as few philosophers would agree to adhere to Carnap's methods, few contemporary philosophers would wish to take on, in their philosophical work, the historical and political burdens that Neurath and Frank carried. Reichenbach once praised logical empiricism in these terms: "The strength of this group lies in its common working program and not in a common doctrine—a program which distinguishes it from philosophical sects, and makes possible progress in research" (Reichenbach 1936, p. 142). It may be that a refusal to engage in political matters is actually a strength of this strand of logical empiricism for it allows a "common working program" that would be undermined by inevitable political disagreements. Supporters of Neurath and Frank would tend to view this judgment as a politically naïve preference for the "icy slopes of logic," or worse, as a sort of tacit support for unsavory political ideals.

5. The End of Logical Empiricism

In the 1960s logical empiricism entered a late and somewhat moderated phase. Many of the more strident methodological pronouncements associated with Schlick, Reichenbach, Carnap, Neurath, and Frank were qualified or combined with new methodological commitments that arguably mark the end of logical empiricism. We see these developments clearly in the work of Carl Hempel. In his early work Hempel sided more or less with Carnap and Reichenbach on the restricted character of philosophical investigations. In his later work he admitted a role for the context of science that marks a break with some of the core ideals of logical empiricism and aligns him more closely with contemporary philosophy.

Hempel is best known for his accounts of confirmation and explanation (Hempel 1965). Two papers along these lines are "Studies in the Logic of Confirmation" and "Studies in the Logic of Explanation". Hempel offers little by way of methodological reflection in the course of carrying out these studies. One reasonable interpretation is that he takes himself to be revealing what confirmation and explanation really are. The tools for these projects are exclusively formal and logical. For example, Hempel distinguished a species of explanation that he dubbed deductive-nomological. A deductive-nomological explanation is one that involves a valid argument with true premises, where at least one premise is a scientific law. Hempel sought to complete this account by further providing formal criteria for a

given statement to be a law. No appeal to the "material" of science such as its subject matter or broader context was permitted.

One can interpret this sort of work as a natural development of Reichenbach's procedure in *Experience and Prediction*. Reichenbach took his job to be to assemble coherent sets of decisions for matching the means and the ends of science. Once these sets of decisions were clear it was possible to see which set, if any, could match the current practice of science. However, Reichenbach seems to have given no special standing to science as it was done. A given set of decisions was not justified because it conformed to current practice. Hempel's formal studies suggest that he *did* take current practice to be more or less justified. All that was required was that one deploy more sophisticated logical tools and demand a level of precision and clarity that ordinary scientists failed to observe. From this methodological perspective, then, the philosopher proposes formal models for a given concept, and tests them by seeing to what degree they validate current scientific practice.

Although it is difficult to know if this is how Hempel conceived of this work at the time it was done, this is certainly how he describes it in his late essay "Valuation and Objectivity in Science" (Hempel 1983).[14] Hempel here uses Carnap's term "explication" and applies it to his earlier work. These explications "propose explicit and precise reconstructions of vague concepts that play an important role in philosophical theories of knowledge," including "confirmation, inductive reasoning, types of explanation, [and] theoretical reduction" (Hempel 1983, p. 379). Hempel emphasizes the extent to which Carnap's own explications depend on an empirical basis. This is because we must first examine how a given concept is deployed in contemporary science before we can propose an adequate, precise substitute for that concept. Carnap had himself granted the need for these preparatory studies. But in line with our earlier discussion of his mature methodology, I would argue that these observations are not responsible for the ultimate justification of Carnap's proposals. For example, Carnap did grant the importance of "our intuitive judgments concerning inductive validity, i.e., concerning inductive rationality of practical decisions (e.g. about bets)" (given at Hempel 1983, p. 380) as reasons to accept his proposals about probability. However, the only real justification for accepting this account of probability is the practical goals that it allows us to achieve. Where we start or what is currently done has no evidential status whatsoever.[15]

For the late Hempel, things are quite different. How we do things now is part of the evidence that we should use to justify our philosophical proposals. This is why appeals to intuitive judgments and empirical observations of scientific procedures are legitimate. Invoking Goodman's celebrated methodology of reflective equilibrium, Hempel insists that explication requires the assumption that

> there is a body of widely *shared* intuitions and unwillingnesses, and that approximate conformity with them provides a justification for acknowledging as sound certain rules of deductive and inductive reasoning. Indeed, without such a body of shared ideas on sound

[14] Wolters 2003 focuses on this later phase of Hempel's development.

[15] Hempel here draws on Carnap's response to Kemeny in Schilpp 1963, p. 978. However, while for Kemeny "our final definition really reproduces the original meaning" (Schilpp 1963, p. 713), in his original discussion of explication Carnap had emphasized the need to dramatically revise meanings in the process of explication (Carnap 1950b, esp. pp. 4, 12–13).

> reasoning, there would be no explicandum, and the question of an explicatory theory could not arise.
>
> (Hempel 1983, p. 380)

Hempel then uses this starting point to relate his earlier formal studies to the "pragmatist" theory of scientific revolutions championed by Kuhn. Hempel takes Kuhn to have shown that scientific practice cannot be accurately reconstructed in formal, logical terms. What is needed, then, is a shift from a restricted, formal rational reconstruction to a "relaxed" rational reconstruction that tolerates some appeal to imprecision: "proper scientific procedures are governed by methodological norms some of which are explicit and precise, while others—including very important ones—are vague" (Hempel 1983, p. 389).

It is but a short step from this methodological proposal of Hempel's to the central methodological quandaries of our own time. For if we take for granted that proposals in the philosophy of science are justified by the current practice of science, then it is tempting to argue that proposals in other areas of philosophy are justified by current practices in other domains, including common sense. Two things are missing from this most recent methodological turn that highlight the gap between the logical empiricists and many contemporary philosophers. First, whatever their considerable methodological differences, the logical empiricists shared Mach's assumption that our current practices stand in need of scientific scrutiny and are ultimately justified only to the extent that they contribute to the ongoing production of genuinely scientific knowledge. Second, the logical empiricists did not hold out any hope of, or desire for, a substantial, autonomous body of philosophical knowledge. This is why the tools they deployed in their critical and constructive projects were no different from the tools deployed by scientists more generally. Logic and empirical observation were thought to be enough to get the job done. Any suggestion that genuine philosophical problems existed that had a depth that required special tools or insights would have been met with a great deal of skepticism. In contemporary philosophy, most of us have overcome that skepticism and I would certainly not urge it in the extreme form found in logical empiricism. At the same time, I take a lasting achievement of logical empiricism to be a healthy reminder that there is always one option available for overcoming the gap between the philosophical tools at our disposal and the philosophical knowledge that we aspire to. We may always elect to give up the quest for that elusive sort of knowledge and content ourselves with what science and other more established pursuits have to offer.

References

Awodey, S. and A. W. Carus (2007). Carnap's Dream: Gödel, Wittgenstein and "Logical Syntax". *Synthese* 159: 23–45.

Ayer, A. J. (ed.) (1959). *Logical Positivism*. New York: Free Press.

Carnap, R. (1932). The Elimination of Metaphysics Through Logical Analysis of Language. Reprinted in Ayer (ed.) (1959), pp. 60–81.

Carnap, R. (1936). Testability and Meaning. *Philosophy of Science* 3: 419–471, 4: 1–40.

Carnap, R. (1950a). Empiricism, Semantics and Ontology. Reprinted in Carnap 1956, pp. 205–21.

Carnap, R. (1950b). *Logical Foundations of Probability*. Chicago: University of Chicago Press.

Carnap, R. (1956). *Meaning and Necessity: A Study in Semantics and Modal Logic*. Chicago: University of Chicago Press.
Carnap, R. (2002). *The Logical Syntax of Language*. A. Smeaton (trans.). Chicago: Open Court.
Carnap, R. (2003). *The Logical Structure of the World and Pseudoproblems in Philosophy*. R. A. George (trans.). Chicago: Open Court.
Cartwright, N., J. Cat, L. Fleck, and T. Uebel (1996). *Otto Neurath: Philosophy Between Science and Politics*. Cambridge: Cambridge University Press.
Chalmers, D., D. Manley, and R. Wasserman (eds.) (2010). *Metametaphysics: New Essays on the Foundations of Ontology*. Oxford: Oxford University Press.
Creath, R. (2007). Quine's Challenge to Carnap. In Friedman and Creath (eds.) (2007), pp. 316–35.
Eklund, M. (2010). Carnap and Ontological Pluralism. In Chalmers, Manley, and Wasserman (eds.) (2010), pp. 130–56.
Frank, P. (1947). The Institute for the Unity of Science: Its Background and Purpose. *Synthese* 6: 160–7.
Frank, P. (1948). The Place of Logic and Metaphysics in the Advancement of Modern Science. *Philosophy of Science* 15: 275–86.
Friedman, M. (1999). *Reconsidering Logical Positivism*. Cambridge: Cambridge University Press.
Friedman, M. (2000). *A Parting of the Ways: Carnap, Cassirer, and Heidegger*. Chicago: Open Court.
Friedman, M. (2007). Coordination, Constitution and Convention: The Evolution of the A Priori in Logical Empiricism. In Richardson and Uebel (eds.) (2007), pp. 91–116.
Friedman, M. and R. Creath (eds.) (2007). *The Cambridge Companion to Carnap*. Cambridge: Cambridge University Press.
Frost-Arnold, G. (2013). *Carnap, Tarski, and Quine at Harvard: Conversations on Logic, Mathematics, and Science*. Chicago: Open Court.
Galavotti, M. C. (2011). On Hans Reichenbach's Inductivism. *Synthese* 181: 95–111.
Hempel, C. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press.
Hempel, C. (1983). Valuation and Objectivity in Science. Reprinted in J. Fetzer (ed.), *The Philosophy of Carl G. Hempel*, New York: Oxford University Press, 2001, pp. 372–95.
Kraut, R. (forthcoming). Three Carnaps on Ontology. In S. Blatti and S. Lapointe (eds.), *Ontology After Carnap*, Oxford: Oxford University Press.
Lavers, G. (forthcoming). Carnap on Abstract and Theoretical Objects. In S. Blatti and S. Lapointe (eds.), *Ontology After Carnap*, Oxford: Oxford University Press.
Mach, E. (1897). *Contributions to the Analysis of Sensations*. La Salle: Open Court.
Mancosu, P. (2008). Tarski, Neurath and Kokoszyńska on the Semantic Conception of Truth. Reprinted in P. Mancosu, *The Adventure of Reason: Interplay between Philosophy of Mathematics and Logic, 1900–1940*. Oxford: Oxford University Press, 2010, pp. 415–39.
Nemeth, E. (2007). Logical Empiricism and the History and Sociology of Science. In Richardson and Uebel (eds.), (2007), pp. 278–302.
Neurath, O. (1932). Protocol Sentences. Reprinted in Ayer (ed.) (1959), pp. 199–208.
Neurath, O. (1935). Unity of Science as Task. Reprinted in O. Neurath, *Philosophical Papers, 1913–1946*, Boston: D. Riedel, 1983, pp. 52–7.
Neurath, O., H. Hahn, and R. Carnap (1929). The Scientific Conception of the World: The Vienna Circle. Reprinted in O. Neurath, *Empiricism and Sociology*, Boston: D. Riedel, 1973, pp. 299–318.

Pincock, C. (2009). Carnap's *Logical Structure of the World*. *Philosophy Compass* 4/6: 951–61.
Price, H. (2010). Metaphysics After Carnap: The Ghost Who Walks? In Chalmers, Manley and Wasserman (eds.) (2010), pp. 320–46.
Psillos, S. (2011). On Reichenbach's Argument for Scientific Realism. *Synthese* 181: 23–40.
Reichenbach, H. (1936). Logistic Empiricism in Germany and the Present State of its Problems. *Journal of Philosophy* 33: 141–60.
Reichenbach, H. (1938). *Experience and Prediction: An Analysis of the Foundations and the Structure of Knowledge*. Chicago: University of Chicago Press.
Reichenbach, H. (1958). *The Philosophy of Space and Time*. New York: Dover.
Reisch, G. (2005). *How the Cold War Transformed Philosophy of Science: To the Icy Slopes of Logic*. Cambridge: Cambridge University Press.
Reisch, G. (2007). From "the Life of the Present" to the "Icy Slopes of Logic": Logical Empiricism, the Unity of Science Movement and the Cold War. In Richardson and Uebel (eds.) (2007), pp. 58–87.
Rescher, N. (2006). The Berlin School of Logical Empiricism and Its Legacy. *Erkenntnis* 64: 281–304.
Richardson, A. (2003). The Scientific World Conception: Logical Positivism. In T. Baldwin (ed.), *The Cambridge History of Philosophy: 1870–1945*, Cambridge: Cambridge University Press, pp. 391–400.
Richardson, A. and T. Uebel (eds.) (2007). *The Cambridge Companion to Logical Empiricism*. Cambridge: Cambridge University Press.
Ricketts, T. (2003). Languages and Calculi. In G. Hardcastle and A. Richardson (eds.), *Logical Empiricism in North America*, Minneapolis: University of Minnesota Press, pp. 257–80.
Russell, B. (1940). *An Inquiry into Meaning and Truth*. London: Allen and Unwin.
Ryle, G. (1949). Discussion of Rudolf Carnap: *Meaning and Necessity*. Reprinted in G. Ryle, *Collected Papers*, vol. 1, London: Hutchinson, pp. 225–35.
Salmon, W. (1991). Hans Reichenbach's Vindication of Induction. *Erkenntnis* 35: 99–122.
Schilpp, P. A. (ed.) (1963). *The Philosophy of Rudolf Carnap*. La Salle, IL: Open Court.
Schlick, M. (1930). The Turning Point in Philosophy. Reprinted in Ayer (ed.) (1959), pp. 53–9.
Schlick, M. (1934). The Foundation of Knowledge. Reprinted in Ayer (ed.) (1959), pp. 209–27.
Schlick, M. (1963). *Space and Time in Contemporary Physics: An Introduction to the Theory of Relativity and Gravitation*. H. Brose (trans.). Mineola: Dover.
Schlick, M. (1985). *General Theory of Knowledge*. A. Blumberg (trans.). Chicago: Open Court.
Stadler, F. (2001). *The Vienna Circle: Studies in the Origins, Development and Influence of Logical Empiricism*. New York: Springer.
Uebel, T. (2007). *Empiricism at the Crossroads: The Vienna Circle's Protocol-Sentence Debate*. Chicago: Open Court.
Uebel, T. (2011). Beyond the Formalist Criterion of Cognitive Significance: Philipp Frank's Later Antimetaphysics. *HOPOS* 1: 47–71.
Uebel, T. (2013). Pragmatics in Carnap and Morris and the Bipartite Metatheory Conception. *Erkenntnis* 78: 523–46.
Van Fraassen, B. (2002). *The Empirical Stance*. New Haven: Yale University Press.
Wittgenstein, L. (1974). *Tractatus Logico-Philosophicus*. D. Pears and B. McGuinness (trans.). London: Routledge and Kegan Paul.
Wolters, G. (2003). Carl Gustav Hempel: Pragmatic Empiricist. In P. Parrini, W. Salmon and M. Salmon (eds.), *Logical Empiricism: Historical and Contemporary Perspectives*, Pittsburgh: University of Pittsburgh Press, pp. 109–22.

CHAPTER 6

ORDINARY LANGUAGE PHILOSOPHY

AVNER BAZ

1. Introduction: Ordinary Language Philosophy, and Its Contemporary Relevance

'Ordinary language philosophy' (henceforth OLP) refers to a general philosophical approach—or more correctly to a family of variously related approaches—that may be situated as follows with respect to its historical protagonists. Ludwig Wittgenstein and J. L. Austin mark the two poles of a sort of magnetic field that may serve to define the approach. Other philosophers whose work may reasonably be said to exemplify OLP—Gilbert Ryle, Peter Strawson, Norman Malcolm, Elizabeth Anscombe, and D. Z. Phillips, among others—could be located somewhere in the field, variously situated with respect to Wittgenstein and to Austin. Then there are other philosophers—G. E. Moore, John Wisdom, Richard Hare, and more recently Philippa Foot, Bernard Williams, Stanley Cavell, Cora Diamond, and Charles Travis, among others—whose relation to the field, while important for an understanding of their work, is more complex. As I note below, OLP may also be seen as having been anticipated by the Socrates of the early Platonic dialogues and by Kant.[1] The aim of this article, however, is not to present the work of any of the above philosophers, but rather to present, clarify, defend, and say something about the contemporary relevance of an approach to the understanding and dissolution of at least very many traditional philosophical difficulties that draws on some of their work.

Naturally, there are many points of contact between this article and Paul Horwich's contribution to chapter 7, which focuses on the later work of Wittgenstein. There is considerable agreement between Horwich's Wittgenstein and my OLP. There are also some

[1] Kierkegaard, in his therapeutic treatment of various confused attempts to express moral, philosophical, and religious positions, is another important precursor of OLP.

important differences. The main points of agreements, as well as the main differences, will be noted in due course.

OLP, as understood in this article, rests on the idea that many (though by no means all) traditional philosophical difficulties arise when we take our words to express thoughts, or to otherwise carry commitments or implications—of the sort, most importantly, that have been taken to generate traditional philosophical difficulties concerning truth, knowledge, meaning, or what have you—in virtue of something called 'their meaning', and irrespective of how we may reasonably be found to mean them, under the circumstances and given our and their history. In relying on the meaning of his words to identify his subject matter well enough and to ensure the sense or intelligibility of what he says, OLP argues, the traditional philosopher[2] expects something of his words that—given the work we ordinarily and normally do by means of them and the conditions under which it may successfully be done—should not be expected of them. He thereby saddles himself with difficulties that derive whatever force they seem to have from that very expectation.

Upon encountering a stretch of philosophical discourse that the ordinary language philosopher (OLPer) suspects of being ultimately nonsensical,[3] or only fit for making sense in ways that would not sustain the philosophical concern supposedly under discussion, one thing the OLPer characteristically does is to appeal to the ordinary and normal use of key words in that stretch of discourse. Contrary to recurrent allegations, the aim of the appeal is not to *prove*, all by itself, that the stretch of discourse makes no sense; for, as I will note below, no such proofs can be had. Rather, the appeal is meant to weaken the hold of the conviction that the philosophical stretch of discourse does and indeed *must* make sense, simply because it consists of familiar words that are put together syntactically correctly. The appeal is also intended to invite those who take that stretch of philosophical discourse to make clear sense to ask themselves what that sense might be, and whether, given the sense or senses it could reasonably be found to have, it really does succeed in expressing a genuine philosophical question or difficulty, or is otherwise fit to do the philosophical work that its author needs or wants it to do.

I should immediately note that there is no better way of clarifying the nature of OLP, and of defending it, than to show how it works—that is, show it *at work*. Unfortunately, however, and for reasons that will become clearer below, it is hard if not impossible to give a *brief* illustration of the practice of OLP and what it can yield. The value of OLP is a function

[2] The term 'traditional philosopher' will be used in this chapter more or less technically, to refer to whoever may be shown to have gotten him or herself into philosophical trouble by forming or otherwise committing him or herself to the above expectation. I do not mean to imply that *all* traditional philosophy is characterized by the above predicament and stands in need of OLP intervention. I will throughout refer to the traditional philosopher as a 'he', because, for any number of reasons—some lamentable, others perhaps not—the tendencies and predicament characteristic of that philosopher have overwhelmingly manifested themselves in men much more than in women. I will use 'she' to refer to the ordinary language philosopher.

[3] The stretch of philosophical discourse need not be put forward assertively by the philosopher. Sometimes it is uttered by the protagonist(s) of what is meant to be an example of a stretch of ordinary (i.e. not philosophically motivated) discourse—an example that is supposed to support the philosopher's argument. In Baz 2012a, I discuss several such examples (or really "examples", for in many cases it is very hard to imagine anyone who is not philosophically motivated uttering the strings of words that the philosopher has put in the mouths of his protagonists).

of the philosophical difficulties it enables us to put to rest: the more pervasive, far reaching, and persistent the difficulty, the greater the value of dissolving it. Simply pointing out *local* and *easily correctable* failures on the part of particular philosophers to make clear sense with their words, or to make the sense they evidently need to make given their broader purposes, would be of little interest, if not merely annoying. But precisely to the extent that the philosophical difficulty is pervasive, far reaching, and persistent, a satisfying OLP dissolution of it would be no simple matter. Later on I will emphasize that the practice of OLP is continuous with our efforts outside philosophy to make sense of other people's and our own words; and just as outside philosophy we cannot rightfully dismiss someone else's *apparently* sensible words as failing to make clear sense unless we first make an effort to see what she might have been trying to say, or needed to say given her broader intentions, so must an OLP diagnosis be guided and informed by an appreciation of the philosophical ambitions, presuppositions, and methodological commitments that have given rise to the difficulty. For this reason, and regretfully, I will not attempt in this article an illustration of the practice of OLP, but will only refer the reader to other texts, others' as well as my own, that illustrate it.

In the history of Western philosophy, the most important predecessor of OLP is Immanuel Kant, especially in the 'Transcendental Dialectic' of the *Critique of Pure Reason*. Kant there argues that when 'empiricists' and 'rationalists' give contrasting answers to questions such as 'Does the world have a beginning in time?' or 'Is the world made of absolutely simple and non-divisible parts?', both sides are attempting to answer a question that rests on a misunderstanding. The competing answers are therefore neither correct nor incorrect, but rather are 'lacking in sense', as Kant puts it (1998, A485/B513). Furthermore, at the root of such traditional 'antinomies' lies the assumption, or fantasy, that our 'categories' and concepts of experience apply to the world 'as it is in itself'—that is, as it is apart from 'the progression of experience', and apart from the 'transcendental' conditions of that experience (cf. Kant 1998, A479/B507 and A493/B521). As a result, Kant says, we *think* we are succeeding in 'employing' our words when we attempt to raise and answer such questions, when we actually are not (Kant 1998, A247/B304). This anticipates Wittgenstein's saying that philosophical problems arise when 'language goes on holiday' (Wittgenstein 1963 (hereafter 'PI'), 38) or is 'like an engine idling' (PI, 132).

Similarly to Kant, OLPers have also found that at the root of any number of traditional philosophical difficulties lies not this or that mistaken answer, or set of mistaken answers, to some perfectly legitimate and intelligible question, but rather, precisely, the question itself, and the assumption that, as raised in the philosophical context (either explicitly or tacitly), it makes clear sense and has a correct answer (see Wittgenstein, 1958, 169; Austin 1964, 4; and Ryle 2000, 22). For both the Kant of the Transcendental Dialectic and the OLPer, what philosophical difficulty calls for in such cases is diagnosis that would make the parties give up their original question as only seemingly intelligible and correctly answerable, as opposed to arbitration that would find which answers to it are correct, and which ones are not (see Kant 1998, A423/B451). For such a diagnosis to be successful it must be informed by a thorough understanding of the presuppositions, processes, and considerations that have brought the philosophical question or difficulty to its present form.

At a high enough level of abstraction, the OLPer agrees with Kant that the philosophical idleness results from the philosopher's attempt to use words apart from certain *conditions*.

Kant likens those sense-conditions to the air resistance that makes it possible for birds to fly, and likens the philosopher to a bird who thinks she could fly (even better) in a vacuum (1998, A5/B8). Wittgenstein likens those conditions to the friction that makes walking possible, and likens the philosopher to someone who thinks he could walk (even better) on slippery ice (PI, 107).

An important difference between Kant and the OLPer is that the former thought it possible to spell out fully and once and for all the conditions for the intelligible employment of each of our philosophically troublesome words, or concepts, whereas the latter thinks that the plasticity, open-endedness, and multi-functionality of language, together with the complexity and ongoing evolution of the human form of life, make *that* impossible. Though useful heuristics and methods could be articulated and exemplified, there is no recipe for what a successful OLP diagnosis of some stretch of philosophical discourse would look like. Each diagnosis must be specifically tailored to the stretch of discourse under consideration, and must begin with an attempt to recover the theoretical commitments, as well as the theoretical ambition, that guide and inform the work responded to.

In attempting to show that, and how exactly, some philosopher has failed to make clear sense with his words, or failed to make the sense he evidently wanted or needed to make, the OLPer must therefore be philosophically informed, but she ultimately relies on nothing more than what we all must rely on when, in the everyday, we try to figure out what sense, if any, someone has made with their words. She relies, namely, on her familiarity with the normal and ordinary uses of the words; on her sense of their potential, when special need arises—philosophical or other—to mean more than, or just differently from, what they normally and ordinarily mean; and on her appreciation of the speaker's or thinker's (philosophical, personal, practical, moral, spiritual …) situation. In other words, the OLPer has no more solid ground for her proposed diagnoses of philosophical difficulties than our more or less shared sense of what makes (what) sense, and under what conditions. But neither do we stand on firmer ground when, outside philosophy, we speak to, and sometimes for, others, and respond to what they say. Here too Kant is relevant, for, in his *Critique of the Power of Judgment*, he argues that the mutual communicability and intelligibility even of empirical judgments ultimately rest, not on concepts or rules, but on a 'common sense (*Sensus Communis*)' that comes to the fore in aesthetic claims.[4]

To see the contemporary relevance of OLP, one only needs to consider the recent debates concerning what is known as *the method of cases*—the method, namely, of theorizing on the basis of the 'application' of terms to theoretically significant cases. Let 'X' stand for some philosophically interesting subject such as knowledge, justification, moral permissibility, causation, necessary truth, intentionality, belief (change of belief), linguistic reference, and so on. As a way either of supporting or motivating some theory of X (or some related subject), or of undermining some theory of X, one central thing analytic philosophers have done is construct (or else simply invoke) cases designed to bring out, either by themselves or in conjunction with other cases, significant features of X; and then they have invited themselves and others to answer questions of the form 'Is this a case of X?', or of the form 'Is this a case of X or of Y?' (where Y is supposed to interestingly

[4] I explore this point of deep affinity between Kant's third *Critique* and OLP in Baz 2015b.

contrast with X[5]). Call any question of this general form, when asked as a way of testing some theory of X, 'the theorist's question'. The general working hypothesis has been that the theorist's question has a correct answer and that good theories of X should fit with the 'correct' answers to the (relevant) theorist's questions.

In recent years, the method of cases has been put under considerable skeptical pressure. One skeptical worry, which originates from Stephen Stich (1988), is that different people might, and do as a matter of empirical fact, disagree with each other in their answers to the theorist's questions; and it is not clear why, or with what right, analytic philosophers should give special weight to what may merely be *their* answers. Another worry, which originates from Robert Cummins (1998), is that there is no way to calibrate the answers we (find ourselves inclined to) give to the theorist's questions—no way of ascertaining that they successfully, let alone reliably, track whatever it is they are supposed to track. These skeptical worries have been instrumental in motivating the recent movement of 'experimental philosophy', and have in turn arguably received some support from the findings of experimental philosophers. It is important to note, however, that analytic philosophers did not need to await those findings in order to have good reason to worry about the soundness of the method of cases as commonly practiced. For, as Ernest Sosa has recently acknowledged, analytic philosophers have fundamentally disagreed (and have known themselves to fundamentally disagree) among themselves in their answers to many of the theorist's questions (2011, 461). Moreover, even where there has been broad agreement among them on the correct answers to *some* questions, those answers have tended to be 'unsystematic', as Gendler and Hawthorne have noted (2005).

In modern Western philosophy in general and in contemporary Analytic Philosophy in particular, reliance on answers to versions of the theorist's question has been central and pervasive. Just how central and pervasive that reliance has been will become evident once it is noted that it may also be found in works that do not appear to deploy the method of cases in its paradigmatic form. Thus, for example, when Grice, in the course of expounding his causal theory of perception, and presumably while looking at the palm of his hand says, 'It looks pink to me' (1989, 234)—taking himself to have thereby succeeded in identifying a particular experience or sense-datum that when properly caused by the presence of the palm of his hand is supposed to constitute, on his theory, his *seeing* that the palm of his hand is pink—he is, in effect, giving an answer to a version of the theorist's question. Every time a philosopher takes it that some cases just are, or are not, cases of X—knowledge, causation, moral permissibility, voluntary action, it looking to one that something is pink, or what have you—irrespective of whatever might lead someone to *count* those cases as cases of X, and irrespective of the *point* or *function* of such counting and of the *conditions* under which the counting may felicitously be carried out, he is relying, in effect, on some particular answer to what I'm calling a theorist's question. The soundness of much philosophical work is at stake, therefore, in the recent debates between the champions of the method of cases and those who have been skeptical about it.

[5] As for example *merely believing* that such and such presumably relates to *knowing* that such and such, or *changing one's belief* that such and such presumably relates to *still believing* that such and such *but in a different way.*

From the perspective of OLP, the fundamental problem, once again, lies not with the particular answers we (find ourselves inclined to) give to the theorist's questions but rather with the assumption, shared by *all* parties to the ongoing debates concerning the method of cases, that the questions themselves are in order—in the simple sense that, as raised by the theorist, they have correct answers[6]—and that, as competent speakers, it ought to be possible for us at least to understand those questions, even if not necessarily answer them correctly. The assumption, in other words, is that the meanings of the theorist's words and how those words are combined (together with the case as described by the theorist) suffice for fixing the theorist's question with a clear enough sense (and a correct answer). So the assumption shared by all parties to the recent debates concerning the method of cases is a version of the assumption that, at the beginning of this article, I identified on behalf of OLP as lying at the very root of any number of traditional philosophical difficulties. I will call it 'the shared assumption' to register the fact that it has been made, not only by proponents of the method of cases, but also by those who have been skeptical of it.

As I note below, the representational-referential and atomistic-compositional conception of language that underwrites the shared assumption, and which tends to be taken for granted in the debates concerning the method of cases, could be challenged *empirically*. Needless to say, empirical theories are underdetermined by their data, which means that no amount of data could prove that prevailing conception of language false, and thereby prove false the shared assumption. I believe it could be shown, however, that the conception has actually been held on the basis of no empirical evidence, and that the evidence we do have actually points in the direction of a Wittgensteinian, pragmatist-holistic conception of language—a conception on which the shared assumption is false.[7]

Another way of questioning the shared assumption would be to practice some OLP. One way of doing *that* would be on a case-by-case basis: take some version of the theorist's question and ask how it might reasonably be understood if raised outside

[6] Even if only relative to the content of someone's or some community's concept of X (see Goldman 2007, 15). The assumption that the theorist's questions are, in principle, clear enough to be answered correctly or incorrectly has been made even by the strongest critics of the method of cases, not to mention its defenders. In pressing his 'calibration' objection to the method, Cummins, for example, never doubts that our answers to the theorist's questions may or may not be 'accurate' (1998, 124). And Jonathan Weinberg, who among the experimentalists is arguably the most skeptical of the method of cases, has recently spoken, in a co-written paper, of the theorist's question as inviting us to 'track philosophical truths' (Weinberg et al. 2010, 332 and 338). The philosophical truths may be relative, or context-sensitive (Ibid., 332); but they are nonetheless truths, which means that the theorist's question is taken to have a correct answer (even if only relatively and context-sensitively). Weinberg's skepticism concerns our ability, and specifically that of philosophers, to track those 'truths' reliably. It does not concern the existence of such truths or the sense of the questions they are supposed to answer. See also Weinberg 2007, which is premised on the assumption that intuitive answers to the theorist's questions are true or false, correct or incorrect. On the other side of the field, those who have responded to Weinberg on behalf of armchair philosophizing have all presupposed—as armchair philosophers themselves have all presupposed—that answers to the theorist's questions are either true or false, correct or incorrect, and may therefore be assessed in terms of their reliability (see, for example, Williamson 2007, Jackson 2011, Ichikawa 2012, Nagel 2012, and Cappelen 2012).

[7] I argue for this in Baz forthcoming a.

philosophy, in the course of everyday experience, with respect to a case such as the one described by the philosopher. One thing that could then emerge is that, depending on the circumstances in which it arises, there are any number of different senses the similarly worded but non-merely-theoretical question could have—different ways the theorist's words would, or could, reasonably be understood, depending on the context in which they are uttered or considered.[8] That would show that, *pace* the shared assumption, the words (and case) by themselves do not suffice for fixing the theorist's question with a determinate sense, and a correct answer. In other words, it would show that the theorist has failed to raise a clear question. That would go some way toward explaining why competent speakers, who by every reasonable criterion mean the word(s) in question in the same way and share the relevant concepts, may nonetheless find themselves disagreeing in their answers to the theorist's question, and why their answers tend to be unsystematic. By themselves, the theorist's words and case fail to provide his audience with sufficient *orientation*—the kind of orientation that is ordinarily and normally provided by a suitable *context*.

In Baz (2012a and 2012b), for example, I consider a fairly broad range of contexts of everyday, non-theoretical encounters with an actual Gettier case, and I invite the reader to see that no question about that case that would naturally arise in *any* of those contexts would be the theorist's intended question—the question, that is, that has elicited the Gettier intuition in many (but not all) people. The theorist's question may be couched *in the same words* in which everyday, non-theoretical questions are couched; but what it *comes to*—what is required for 'understanding' it and answering it 'correctly'—is altogether unlike what the similarly worded questions that arise naturally in the course of everyday experience come to. On the basis of those OLP reminders, I then argue that whatever the theorist's question invites us to do, it is *not* something that we regularly have to do as part of our everyday employment of our words; and I argue that it is therefore unclear what, if anything, is revealed by the answers we (find ourselves inclined to) give to that question.

Another way of questioning the shared assumption by way of OLP would be to tackle the conception(s) of language that might be thought to support it. This is what I take Wittgenstein to be doing in his rule-following remarks (and in much of the rest of the *Investigations*). If the meaning of a word is thought of as something like a rule that determines in advance its sense or proper understanding, or the contribution it makes to the overall sense of utterances in which it features—and it seems to me that those who take the theorist's question to have a determinate sense and a correct answer must so think of the meanings of the words that make it up—then Wittgenstein's remarks invite us to see that those who thus think of the meanings of words expect of their words, or of their meanings, something they cannot provide. Indeed, what Wittgenstein's remarks show is that we do not really know what it could possibly mean for our words, or their meanings, to determine all by themselves what contribution they (may) make to the overall sense or proper understanding of utterances in *every* context, let alone apart from *any* particular context. So we do not know what it could possibly mean for the shared assumption to be true.

[8] Compare Travis 1991, 250.

2. Elucidating Key Points, Averting Misunderstandings

Having given the above rough characterization of OLP (as presented and defended in this chapter), I now turn to emphasize and/or elucidate a number of points that are crucial to a proper understanding of it. Some of the most common objections to OLP rest on a failure to appreciate or understand one or more of these points. Alternatively, some of those objections may be valid, but only when raised against versions of OLP different from the one presented and defended here. I have ordered the issues, not by importance, but rather in such a way that, as much as possible, those that come before prepare the stage for those that come after.

1. It has often been argued against (this or that version of) OLP that it focuses merely on *words*, as contrasted with the things or worldly phenomena that philosophers have traditionally tried to become clearer about—knowledge, freedom, truth, meaning, justice, and so on (see Horwich, 2013, 19). And in recent years, analytic philosophers such as Scott Soames (2003), Ernest Sosa (2007), Timothy Williamson (2007 and forthcoming), and Jason Stanley (2008) have made considerable efforts to undo the so-called 'linguistic turn' in philosophy and to convince us that, *pace* OLP as they understand it, it is not true that philosophical problems are all at bottom linguistic problems. What philosophers have mostly been interested in, they have insisted, are *things and their natures*, not words and their meanings. Thus, for example, Williamson has recently contended that, *pace* OLP as he understands it, 'the epistemologists' underlying object is knowing itself, not the verb "to know" or the concept of knowing' (Williamson forthcoming). This critique of OLP is doubly misguided. First, OLP's appeal to the ordinary and normal use of philosophically troublesome words is an invitation to remind ourselves of our linguistic *practices*, and of the humanly significant *situations* and *worldly conditions* in which philosophically troublesome words normally and ordinarily do their work. Far from focusing merely on words and neglecting the 'extra-linguistic' world, OLP seeks to lead us back, not just to the world we speak *of*, but also, and primarily, to the world we speak *in*.[9] And second, to the extent that some particular OLP diagnosis is successful,

[9] Horwich's response in this volume to the charge that Wittgenstein's work focuses on *language* rather than *reality* strikes me as still mistakenly beholden to a representationalist-referential (even if also deflationist) conception of language and linguistic meaning. When he contends that philosophers, on Wittgenstein's view, are not really mistaken about language but rather are mistaken 'about the world itself', which he then glosses in terms of the 'objects' and 'properties' our words refer to (p. 143), he fails to take to heart, it seems to me, Wittgenstein's questioning of the 'Augustinian' assumption that the meaning of words—in particular, philosophically troublesome words such as 'know', 'true', and 'cause'—is best understood on the model of 'object and designation' (PI, 293), his urging us to recognize the variety (and complexity) of ways in which words are used (see PI, 182), and his proposal that at least certain types of traditional philosophical difficulties will not go away until we 'make a radical break with the idea that language always functions in one way, always serves the same purpose: to convey thoughts—which may be about houses, pains, good and evil, or anything else you please' (PI, 304).

it shows that it is actually the traditional philosopher who has ultimately offered us nothing but words—'air-structures', as Wittgenstein famously puts it (PI, 118). Moreover, all too often it is precisely the idea that it should be possible for us to study philosophically interesting 'objects' such as knowledge, truth, understanding, and so on, *directly*—that is, not by way of studying the ordinary and normal use of the words by means of which we refer to (express, invoke, appeal to, claim …) those 'objects'—that leads us to philosophical emptiness. As I will note, this philosophically fateful idea rests, in turn, on a conception of language that is both philosophically and empirically challengeable.

2. OLP, as I have presented it, may seem purely negative, or anyway purely critical, providing no positive understanding of its own of anything—its value being wholly a function of the pervasiveness, persistence, and depth of the philosophical difficulties it enables us to put to rest (to the extent that it does). There is some truth in this. OLP, as I note below, puts forward no *theses* or *theories* (see PI, 109 and 128). It is essential to it that it teaches us nothing *new*. The understanding it offers is not of *that* kind. At the same time, however, it offers us something whose value goes beyond whatever philosophical headaches it enables us to alleviate. In appealing, in the face of philosophical difficulties, to our sense of what makes (what) sense and under what conditions, it reminds us of aspects of our life and world: things that we must at some level already know—for they are part of what we know in knowing how to use competently the words under consideration, and to respond competently to other people's use of them—but which we have not been able to see clearly, or notice, in part because theoretical commitments and ambitions have obscured our vision of them, and in part precisely because of their great familiarity (see PI, 129). We need to be brought to *re-turn* to those aspects and features of our life with words in order to *see* them clearly.
3. I said that the OLPer characteristically appeals to the ordinary and normal *use* of the philosophically troublesome word that is under consideration, as a way of becoming clearer about its meaning and thereby finding out what, if anything, the philosopher could mean by it. All too often, however, detractors of OLP have taken the question of use to be the essentially *empirical* (statistical) question of what words people tend to utter under certain (types of) objectively defined circumstances.[10] One obvious problem with the OLPer's appeal to ordinary and normal use, *thus* understood, is that it seems woefully overly generalized—relying as it does on nothing more than the OLPer's own linguistic tendencies (assuming she can even get *those* right from her armchair). Another apparent problem, which I discuss in #11, is that the question of use, understood as an *empirical* question, seems to be very different from the question of sense, so it is hard to see how one could go directly, as the OLPer allegedly does, from answers to the first question to answers to the second.

   This understanding of the OLPer's 'use' is a serious misunderstanding, and the two problems I have mentioned are only pseudo-problems. The relevant notion

[10] A very clear example of this is Soames' understanding of 'use' in his critique of OLP (cf. Soames 2003, 129).

of 'use' here is not empirical but, in a sense, *normative*. It refers to a certain kind of human *achievement*—however humble and everyday—one that contrasts not with mentioning the words, but with letting them *idle*, or failing to do any (real) work with them.[11] And this means that whether certain uttered words are actually being *used* on the occasion of their utterance, inside or outside philosophy, and if so how, is never a straightforward empirical matter (which does not mean that empirical data could have no bearing on that question). The OLPer's appeal to use is best understood, not as an *assertion* of empirical *fact* that is put forth on the basis of very slim evidence, but rather as an *invitation* to her audience, and first and foremost to the traditional philosopher himself, to see whether they share her sense of what makes (what) sense and under what conditions.[12] As Kant suggests, and for reasons I will elaborate, the question of sense is, at bottom, *akin to* paradigmatic aesthetic questions.[13] At the same time, however, since both the traditional philosopher and the OLPer are presumably competent employers of the words in question who, outside philosophy, would presumably be able to communicate with each other smoothly and effectively by means of those words, there is good reason to suppose that they do agree, at least to a considerable degree, in their sense of what sense can be made by means of those words, and under what conditions.

4. The OLPer's characteristic appeal is *not*, ultimately, to anything aptly called 'rules' (of use or usage, of sense or meaning, or of anything else); and the philosopher is not, ultimately, charged with a violation of rules (Contra Horwich 2012, 184; and contra Baker and Hacker's (1980, 1985) influential reading of Wittgenstein). On the understanding of language that informs the practice of OLP as I understand it, our use of words 'is not everywhere circumscribed by rules', as Wittgenstein puts it (PI, 68); and even this way of putting things is really too weak and potentially misleading, because, as already noted, the upshot of Wittgenstein's remarks on 'rule-following' is that we do not really know what it could possibly mean for our use of words to be everywhere circumscribed by rules—in such a way that creative and yet mutually intelligible extensions of their use would be neither possible nor sometimes called for, and truth and falsity, as well as sense and nonsense, would not ultimately depend on nothing more (nor less) solid than our 'agreement in form of life' or 'in judgments' (PI, 241 and 242). For this reason, the claim that our use of words is not everywhere circumscribed by rules is *not* a *theoretical* claim, for it is not the negation of a real, intelligible alternative (more on this in point 12). As

[11] The Austinian inflection of Wittgenstein's concern with what he calls 'use' is the emphasis, in Austin 1999, on what we *do* with our words, and Austin's reminders to the effect that there are *conditions* for doing one thing or another with one's words. You can no more just decide, or bring it about just by willing, that your utterance will constitute, for example, a *claim*, than you could just decide or bring it about by willing that your saying 'I'm sorry' with no one to hear would constitute an *apology*.

[12] This understanding of the force of the philosophical appeal to ordinary language is elaborated in the opening pages of Cavell (1979).

[13] See Cavell (1969, 73–96). I pursue this link between Kant's account of beauty and the practice of OLP in Baz (forthcoming b).

Stanley Cavell has argued in his presentation of Wittgenstein's 'vision of language', an individual who did not possess, to some degree, the ability to 'project words' more or less creatively into new contexts would never become a competent speaker of a natural language (1979). Cavell has also argued, following Wittgenstein, that our ability to communicate effectively by means of language rests on nothing more (nor less) solid than 'our sharing routes of interest and feeling, modes of response, senses of humor and of significance and of fulfillment, of what is outrageous, of what is similar to what else, what a rebuke, what forgiveness, of when an utterance is an assertion, when an appeal, when an explanation—all the whirl of organism Wittgenstein calls "forms of life" '. (1969, 52). Since language is plastic and always, in principle, open for more or less creative extensions, and since it is ultimately grounded in our sense of what makes (what) sense, rules are at most more or less useful generalizations that hold, at best, only for the most part, and for now. (Such generalizations, I would also argue, are all that contemporary semantic theorizing could plausibly claim to discover; and it is no wonder that it is virtually always possible to come up with counterexamples to any semantic rule the theorist proposes (as Horwich notes, 2012, 4)).

5. Since language is plastic and not everywhere circumscribed by rules, the mere fact that the philosopher appears to be using his words differently from how we ordinarily and normally use them does not by itself show that he is not making sense, or not making the sense that he evidently wants or needs to be making. The philosopher *could* be making new sense with his words. Whether or not he has done so successfully is something that needs to be *made out* (contra Horwich, 144); but the following points must be borne in mind:
   a. Typically, the philosopher does not take or present himself as making new sense with his words, or as meaning something else by them than what we ordinarily and normally mean by them. In proposing an account of knowledge, truth, meaning, and so on, the philosopher typically takes himself to be offering an account of what *we* normally and ordinarily call 'knowledge', 'truth', 'meaning', and so on. As Cappelen has recently put it, when philosophers are interested in knowledge, for example, 'they are interested in the phenomenon ordinary speakers of English talk about when they say things like "John knows that Samantha is in Paris" ' (2012, 27).
   b. If the philosopher did (successfully) try to make new sense with his words or mean them differently from how we normally and ordinarily mean them outside philosophy, he would invite upon himself the question of *relevance*: why should we care about whatever it is that the philosopher means by 'knowledge', or 'truth', or 'meaning', and so on? (I do not say that the philosopher could not possibly have a good answer to the question of relevance; I only say that he would owe us one.)
   c. What the philosopher cannot legitimately or coherently do—and what he all too often is in effect trying to do—is take and present himself as meaning by his words what we normally and ordinarily mean by them outside philosophy, while evidently uttering them apart from any of the sorts of contexts in which they normally and ordinarily are used (see PI, 117). For again, how words may be meant is a matter of how they may reasonably be taken (to have been meant), under the

circumstances and given their and their utterer's history; and the philosopher cannot enact those sense-conditions by fiat, or by a particular concentration of his mind.

6. The previous point applies to the OLPer just as much as it applies to the philosopher(s) she criticizes. The OLPer does not, or anyway should not, take herself to be immune to the danger—to which all speakers are vulnerable—of getting lost with one's words.
7. Though OLP takes it that certain tendencies and aspirations characteristic of traditional philosophy naturally lead us to get lost with our words, it does not claim that *all* philosophy (whatever *that* might refer to) is nonsense, and therefore is *not* self-undermining (compare Horwich, 142). OLP (as here understood and defended) does not aim—surely hopelessly—to bring all philosophizing to an end; its aim—and herein lies a deep affinity between it and certain forms of existentialism—is to get the philosopher (in each of us) to remember that he or she is not absolved of the conditions of sense, and cannot intelligibly speak, or mean his or her words, eternally and from nowhere, so to speak—altogether outside of time and place.
8. It has often been said, by both detractors and proponents of OLP, that it typically charges the traditional philosopher with the production of *nonsense* (or meaninglessness) (see Williamson 2013, e10). This common idea is not exactly wrong; and it is true that Wittgenstein tends to present himself as aiming to expose philosophically induced nonsense (*Unsinn*) (cf. PI, 464 and 524). I have come to think, however, that the OLPer's use of 'nonsense' to name her target is problematic, and potentially misleading. The failure to make clear sense with one's words, or to make the sense one evidently wishes or needs to make given one's theoretical ambitions and commitments, can take any number of forms. There are indefinitely many ways of getting lost with our words. For this reason, as I said, OLP's diagnoses—though informed by a certain broad understanding of the forces that lead philosophers to get lost with their words—need to be specifically tailored to the work, or works, they respond to. And this means that the charge of 'nonsense' is likely to come either too early or too late: until she has managed to produce a successful diagnosis, the OLPer has not entitled herself to dismissing some stretch of philosophical discourse as 'nonsensical'; insofar as she *has* produced a successful diagnosis, calling the diagnosed work(s) 'nonsensical' would not be likely to capture well what it has shown. Moreover, the blanket charge of 'nonsense' would be apt to antagonize the philosopher(s) to whom the diagnosis should first and foremost be addressed.[14]
9. Since there are no rules that could, by themselves, decide what sense, if any, some stretch of human discourse makes, the final court of appeal is always, ultimately, competent speakers' sense of what makes (what) sense. We—that is, all competent speakers of the language—are, in principle, equal authorities when it comes to sense.

[14] An exemplary OLP text in this respect is Stanley Cavell's 'Knowing and Acknowledging' (in Cavell 1969). Responding to a *particular form* of debate between a skeptic and an anti-skeptic about other minds, Cavell nowhere charges either of the two parties with the production of sheer nonsense. His terms of criticism are far more specific than that.

The OLPer's aim must therefore be to offer the sort of diagnosis that would, *ideally*, be accepted by the author(s) of the work(s) diagnosed.[15] In this and other respects, OLP's diagnoses are similar to the interventions of the Socrates of the early Platonic dialogues. At least on one (Kierkegaardian) reading of those dialogues, Socrates has successfully made his point when he has gotten his interlocutors to acknowledge that they don't know what *they* have in mind, because *their* words—the very words with which they have attempted to justify fateful actions or decisions, or to account for what they themselves claim to be most important—'go around and refuse to stay put' for *them*, as Euthyphro puts it to Socrates (Plato 1975, 11).

I said that the OLPer's diagnosis should *ideally* be accepted by its target, because there are obviously great psychological forces that make it unlikely that the philosopher him- or herself would accept the diagnosis, however clear, sensitive, and accurate it might be. Note, however, that in everyday discourse, getting someone to see, and acknowledge, that he has not (yet) said anything clear, or that what he said can only reasonably be understood in ways that do not fit with his or her broader intentions and commitments, is a common occurrence. In this way, the practice of OLP is continuous with everyday practice, and rests on the same conditions.[16]

10. Analogously to the realm of aesthetics, the fact that we must *ultimately* fall back on our sense of what makes (what) sense does not mean that anything goes in the realm of linguistic sense and that there is no room in that realm for rational considerations, and failure. More specifically, both inside and outside philosophy the *apparent* sense or intelligibility of a certain stretch of discourse, and its author's original sense that it makes clear sense (see Williamson forthcoming), are not by themselves good evidence that it does make clear sense, for those may be accounted for by the fact that the words themselves are (typically) perfectly familiar and are put together syntactically correctly. Furthermore, in some cases, the philosopher's words are such that we could easily enough imagine contexts in which they *would* make sense; and it's just that the philosopher evidently (and typically fully consciously) utters them and expects them to make clear sense *apart from* any of those contexts (see PI, 117, and Wittgenstein 1969, remark 10). Moreover, as both Kant and Wittgenstein have stressed, the apparent sense of some stretch of philosophical discourse is often sustained by *pictures*—of the soul or mind as a gaseous substance separable from and hidden inside a person's body, for example, or of knowledge as a super-strong connection between a mind and a fact, or of the meaning of a

[15] Compare Cavell: '[A] test of his [the OLPer's] criticism must be whether those to whom it is directed accept its truth, since they are as authoritative as he is in evaluating the data upon which it will be based … But what it means is not that the critic and his opposition must come to *agree* about certain propositions which until now they have disagreed about …What this critic wants, or needs, is a possession of data and descriptions and diagnoses so clear and common that apart from them neither agreement nor disagreement would be possible' (1969, 241).

[16] Interestingly, recent attempts to defend the philosophical 'method of cases' as commonly practiced have rested on the claim that *that* practice is continuous with the everyday, nonphilosophical employment of our words (see Williamson 2004 and 2007, and Cappelen 2012). For reasons stated in section 1.1, I think *this* claim to continuity is mistaken.

word as rails that stretch to infinity and determine its sense or proper use in all possible contexts, or of the world as a whole as an object that may come in and out of existence at a certain point in time, and so on. What the OLPer's diagnoses (when successful) reveal is that pictures may create the illusion of sense, but cannot ensure sense (cf. Kant 1998, A485/B513 and PI, 422–5). So it would not do, once Kant, for example, has exposed the problematic nature of the questions that have generated the antinomies, for the empiricist and the rationalist simply to respond that the questions are apparently intelligible and make sense to *them*. They would need, at the very least, to find ways of legitimately rejecting Kant's transcendental idealism—in light of which those questions are rendered problematic. And they'll also need to account for the fact that they have given contrasting answers to those questions and that it's none too clear how the apparent disagreement between them could possibly be settled.

11. A common allegation against OLP is that its procedures fundamentally ignore the distinction between 'semantics' and 'pragmatics'. The OLPer notes, it is commonly alleged, that some particular stretch of philosophical discourse is, or seems, or sounds, odd, unnatural, nothing like the sort of thing that might naturally be said outside philosophy, in the course of everyday discourse; and from this she (allegedly) concludes that the philosopher's words make no sense, or that he is not making sense by means of them. But, the complaint continues, as Grice (1989) and Searle (1999) have taught us to recognize, the oddness or seeming unnaturalness of the philosophical stretch of discourse is a poor basis for dismissing it as nonsensical, for it may be that the philosopher is saying something that makes perfect sense, something that may well even be true, and it's only the *speech-act of uttering it* that would be odd or unnatural or misleading or otherwise problematic *outside philosophy*, or apart from particular circumstances that evidently do not hold in the philosopher's context.

    This line of objection to OLP, which one hears everywhere in one version or another (most recently in Williamson forthcoming) misconstrues it, and begs the question against it. Both Grice and Searle give a false and misleading account of the point of departure of OLP. Both suggest that the ordinary language philosopher *begins* by 'noticing' or 'observing' that a particular form of words that the philosopher has produced would be 'odd' or 'inappropriate' or 'bizarre' to utter under normal circumstances, or apart from some special circumstances (Grice 1989, 3 and 235; Searle 1999, 141–2). Grice and Searle would have us think that what would and would not be appropriate to say under, or apart from, this or that set of circumstances is *all* the ordinary language philosopher has got to go on in her criticism of the philosopher's words. And then they offer their counter-explanation of the ordinary language philosopher's alleged data: What the philosopher says is perfectly clear and, in particular, is either true or false (valid or invalid, sound or unsound); it's just that actually saying it apart from suitable circumstances—where, again, it is assumed that there is no question about the identity of the *it* that would be said, the 'proposition' or 'thought' that would be 'expressed'—would somehow be inappropriate or misleading or otherwise conversationally infelicitous. To support their counter-explanation, they remind us that the utterance of *any* English sentence—even a sentence as simple and (presumably) impossible not to understand as 'This

pillar-box is red' (Grice 1989, 235) or 'He has five fingers on his left hand' (Searle 1999, 143)—would be odd if made in circumstances in which we could find no point for it. Even so, they insist, it would still be clear *what* the utterer was saying, even if not *why* he said it; and what he would say could very well still be *true*. In fact, its being obviously or trivially true may be precisely the reason why we find saying it odd.

All of this might have been pertinent for an assessment of OLP, if the ordinary language philosopher really began where Grice and Searle say she begins. But she doesn't. She doesn't merely find the philosopher's stretch of discourse odd, or bizarre, or out of place in ordinary contexts, and she certainly does not find it obviously or trivially true. Rather, she finds that, as produced by the philosopher, it is lacking in clear sense, or only fit for being made sense of in ways that would actually undermine the philosopher's project. And insofar as she is able to show this, the Gricean machinery of conversational 'implicature' and Searle's 'assertion fallacy' are both wholly beside the point.

'The Issue', Travis has argued against Grice, 'is one of *making sense*; not one of what we wouldn't say' (1991: 241). Travis then goes on to show, masterfully and convincingly it seems to me, that none of the English words or locutions that Grice himself is relying on in presenting his theory of saying, meaning, and 'implicature' (precisely the theory that was supposed to allow us to legitimately set aside the objections that OLPers have raised against traditional philosophizing), and in presenting his 'causal theory of perception', is fit to do the theoretical work Grice evidently needs it to do, given his theoretical commitments and ambitions.

12. The practice of OLP is informed by a certain understanding of the nature of (certain kinds of) philosophical difficulties, but it is not based on anything aptly called 'a theory' (of language, or meaning, or anything else). Contra Williamson (2013, e10), offering the kind of diagnosis that would ideally get the producer of a stretch of discourse, or at least his audience, to see that he has not (yet) succeeded in saying something clear, or that he does not really know what he is saying or thinking, need not rely on *any* 'assumptions about the nature of meaning'. That we can sometimes get lost with our words, and that appeals to the ordinary and normal use of our words can help us find our way again, or find that we have no clear way, is part of normal practice, part of what *any* theory of language would need to explain. The 'assumption' that the traditional philosopher is not immune to the risk faced by all speakers of getting lost with their words is hardly *theoretical*; and neither is the idea that relying on the meanings of one's words to ensure the sense of what one is saying, while at the same time uttering them apart from any of the contexts in which they normally and ordinarily do their works—as philosophers have been prone to do—increases that risk significantly.
13. Nor, and here I am in full agreement with Horwich, is Wittgenstein offering a *theory* of meaning when he famously suggests that 'for a *large* class of cases—though not for all—in which we employ the word "meaning", one could explain it thus: the meaning of a word is its use in the language' (PI 43). And since he is not offering a theory, he is not, preposterously, offering one 'on almost no

evidence' (Williamson 2013, e8). As Williamson himself has usefully reminded us, the ability to use a word competently in a wide enough range of contexts, and to respond competently to other people's uses of it, is our *ordinary criterion* for 'knowing the meaning' of that word (Williamson 2007, 97, and 2005, 11–12). And this, again, means that the OLPer's invitation to remind ourselves of how the philosophically troublesome word is ordinarily and normally used, as a way of becoming clearer about what may be said by means of it, rests on no substantive theoretical commitments. By contrast, the traditional philosopher's assumption that the sense of some stretch of philosophical discourse is, in principle, ensured by whatever it is that the philosopher's words carry with them from one occasion of use to another—call it 'their meanings'—*is* a substantive theoretical assumption that may be challenged, both philosophically and empirically.

14. Though the practice of OLP rests on no substantive theoretical presuppositions, empirical studies and observations concerning language use and acquisitions *could* play a role in the attempt to defend the practice and to weaken the hold of the theoretical commitments that have tended to inform both the assumption that the traditional philosopher is—even must be—making clear sense with his words and the resistance to OLP's diagnostic interventions. Wittgenstein famously urges the philosopher not to *think*—about what *must* be true of all of the things we call 'games', or 'languages', and more broadly about what *must* be involved in the acquisition and use of language in general and philosophically troublesome words in particular—but rather to *look and see* (PI 66).[17] I believe it can be shown that though contemporary analytic philosophers tend to present themselves as open and attentive to the findings of empirical science, they routinely rely on representational-referential (as opposed to broadly pragmatist) and atomistic-compositional (as opposed to broadly holistic) assumptions concerning language use and acquisition—not only in their rejection of OLP but also in their presentation and defense of their own practice—that are actually supported by no empirical evidence.[18] Coming back to point 1, the 'objects' ('things', 'items') to which philosophically troublesome words such as 'know', 'cause', 'mean', 'pain', and so on, are supposed to refer, and which are supposed to be theoretically separable from the meanings of those words (or from the concepts they embody)—as those meanings (or concepts) manifest themselves

[17] One thing Wittgenstein says we'll see if we look without theoretical prejudice is that among the various things we call or refer to by a common name such as 'game' or 'language' there holds a 'family resemblance'; and this as against the prevailing philosophical assumption that there *must* be something in common to all of the things we call by the same word, which (if we are competent) *makes us* call or refer to them by that word. This claim of Wittgenstein's has received significant empirical support (see Rosch & Mervis 1975, Rosch 1978). In Baz (forthcoming a) I propose, on the basis of empirical studies of first language acquisition, that the idea of 'family resemblance' may be extended in a non-representationalist direction, to describe not just the relation among the different *things* referred to by the same word but also more broadly the relation among the different *functions* of philosophically troublesome words such as 'know' and 'cause'.

[18] I argue for this in detail, on the basis of empirical studies of first language acquisition, in Baz forthcoming a.

in our ordinary and normal use of those words—those 'objects' are, at best, posits of what may turn out to be a bad theory of language. At worst, they are nothing more than shadows cast by the ways we talk, to use Huw Price's apt image (2011, 319).[19]

References

Austin, J. L. (1964). *Sense and Sensibilia*. New York: Oxford University Press.
Austin, J. L. (1999). *How to Do Things with Words*. Cambridge, MA: Harvard University Press.
Baker, G. P. and Hacker, P. M. S. (1980). *Wittgenstein: Understanding and Meaning*. Oxford: Blackwell.
Baker, G. P. and Hacker, P. M. S. (1985). *Wittgenstein: Rules, Grammar, and Necessity*. Oxford: Blackwell.
Baz, A. (2012a). *When Words are Called For*. Cambridge, MA: Harvard University Press.
Baz, A. (2012b). 'Must Philosophers Rely on Intuitions?'. *Journal of Philosophy* 109: 316–37.
Baz, A. (forthcoming a). 'On Going (and Getting) Nowhere with Our Words: New Skepticism about the Method of Cases'. *Philosophical Psychology*.
Baz, A. (forthcoming b). 'The Sound of Bedrock: Lines of Grammar Between Kant, Wittgenstein, and Cavell'. *European Journal of Philosophy*.
Cappelen, H. (2012). *Philosophy Without Intuitions*. New York: Oxford University Press.
Cavell, S. (1969). *Must We Mean What We Say?* New York: Cambridge University Press.
Cavell, S. (1979). *The Claim of Reason*. New York: Oxford University Press.
Cummins, R. (1998). 'Reflection on Reflective Equilibrium'. In DePaul and Ramsey, *Rethinking Intuitions: The Psychology of Intuition and Its Role in Philosophical Inquiry*. Oxford: Rowman and Littlefield, 113–28.
Gendler, T. and Hawthorne, J. (2005). 'The Real Guide to Fake Barns: A Catalogue of Gifts for your Epistemic Enemies'. *Philosophical Studies* 124: 331–52.
Goldman, A. (2007). 'Philosophical Intuitions: Their Target, Their Source, and Their Epistemic Status'. *Grazer Philosophische Studien* 4: 1–26.
Grice, P. (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.
Horwich. P. (2012). *Wittgenstein's Metaphilosophy*. Oxford: Clarendon Press.
Horwich, P. (2013). 'Reply to Timothy Williamson's Review of *Wittgenstein's Metaphilosophy*'. *European Journal of Philosophy* 21 (issue supplement): 18–26.
Jackson, F. (2011). 'On Gettier Holdouts'. *Mind and Language* 26: 468–81.

[19] The recurrent idea that philosophers should study philosophically interesting "objects" directly, rather than by way of studying our concepts of those "objects"—as those concepts manifest themselves in our use of the corresponding words—is essentially the same as the position Price (2011) calls 'object naturalism'. About object naturalism Price says that it 'rests on substantial theoretical assumptions about what we humans do with language—roughly, the [representationalist] assumption that substantial "word-world" semantic relations are part of the best scientific account of our use of the relevant terms' (2011, 190). Price argues, however, that 'by the naturalist's own lights, the [representationalist's] proto-theory ought to count as an hypothesis about what it is right to say about language itself, from a naturalistic standpoint. If it turned out to be a bad hypothesis—if better science showed that the proto-theory was a poor theory—then the motivation for the Naturalist's version of the matching game [of words and worldly items] would be undermined' (2011, 5; see also 209; see also Price 2013, 4, 14, and 25).

Ichikawa, J. (2012). 'Experimentalist Pressure Against Traditional Methodology'. *Philosophical Psychology* 25: 743–65.
Kant, I. (1998). *Critique of Pure Reason*. Guyer, P. and Wood, A. (eds. and trans.). New York: Cambridge University Press.
Margolis, E. and Laurence, S. (1999), *Concepts*. Cambridge, MA: MIT Press.
Nagel, J. (2012). 'Intuitions and Experiments: A Defense of the Case Method in Epistemology'. *Philosophy and Phenomenological Research* 85: 495–527.
Plato (1975). *The Trial and Death of Socrates*. Grube, G. M. A. (tr.). Indianapolis: Hackett.
Price, H. (2011). *Naturalism Without Mirrors*. New York: Oxford University Press.
Price, H. (2013). *Expressivism, Pragmatism, and Representationalism*. New York: Cambridge University Press.
Rosch, E. and Mervis C. (1975). 'Family Resemblances: Studies in the Internal Structure of Categories'. *Cognitive Psychology* 7: 573–605.
Rosch, E. (1978). 'Principles of Categorization'. In *Cognition and Categorization*. Rosch, E. and Lloyd, B. B. (eds.). Hillsdale, NJ: Erlbaum, 27–48. Reprinted in Margolis and Laurence 1999.
Ryle, G. (2000). *The Concept of Mind*. Chicago: Chicago University Press.
Searle, J. (1999). *Speech Acts*. New York: Cambridge University Press.
Soames, S. (2003). *Philosophical Analysis in the 20th Century, Volume 2: The Age of Meaning*. Princeton: Princeton University Press.
Sosa, E. (2007), 'Intuitions: Their Nature and Epistemic Efficacy'. *Grazer Philosophische Studien* 74: 51–67.
Sosa, E. (2011). 'Can there be a Discipline of Philosophy and Can it be Founded on Intuitions?'. *Mind and Language* 26: 453–67.
Stanley, J. (2008). 'Philosophy of Language in the 20th Century'. In *Routledge Guide to 20th Century Philosophy*. M. Dermont (ed.). New York: Routledge, 382–437.
Stich, S. (1988). 'Reflective Equilibrium, Analytic Epistemology, and the Problem of Cognitive Diversity'. *Synthese* 74: 391–413.
Travis, C. (1991). 'Annals of Analysis'. *Mind* 100: 237–64.
Weinberg, J. (2007). 'How to Challenge Intuitions Empirically Without Risking Skepticism'. *Midwest Studies in Philosophy* 31: 318–43.
Weinberg, J., Gonnerman, C., Buckner, C., and Alexander, J. (2010). 'Are Philosophers Experts Intuiters?'. *Philosophical Psychology* 23: 331–55.
Williamson, T. (2004). 'Philosophical "Intuitions" and Skepticism about Judgment'. *Dialectica* 58: 109–53.
Williamson, T. (2005). 'Armchair Philosophy, Metaphysical Modality, and Counterfactual Thinking'. *Proceedings of the Aristotelian Society* 105: 1–23.
Williamson, T. (2007). *Philosophy of Philosophy*. New York: Oxford University Press.
Williamson, T. (2013). 'Review of *Wittgenstein's Metaphilosophy* by Paul Horwich'. *European Journal of Philosophy* 21 (issue supplement S2): e7–e10.
Williamson, T. (forthcoming). 'How Did We Get Here from There: The Transformation of Analytic Philosophy'. *Belgrade Philosophical Annual*.
Wittgenstein, L. (1958). *The Blue and Brown Books*. Oxford: Blackwell.
Wittgenstein, L. (1963). *Philosophical Investigations*. G. E. M. Anscombe (trans.). Oxford: Basil Blackwell.
Wittgenstein, L. (1969). *On Certainty*. Anscombe, G. E. M. and von Wright, G. H. (eds.). Paul, D. and Anscombe, G. E. M. (trans.). New York: Harper and Row.

CHAPTER 7

WITTGENSTEIN'S GLOBAL DEFLATIONISM

PAUL HORWICH

1. Introduction

Ludwig Wittgenstein's mature philosophy of philosophy is *critical* rather than *constructive*. He makes no attempt to devise a system of a priori principles to explain fundamental aspects of language, the mind, reality, etc. On the contrary, his goal is to undermine the very idea that such theorizing is a reasonable endeavor.[1]

That critique isn't directed at philosophy *in general*. He fully appreciates that there are various different but loosely related kinds of intellectual activity to which the label "philosophy" is often applied by their adherents—with none of them having exclusive rights to the word. And his objection is merely to a specific member of this cluster. It targets a certain traditionally dominant form of philosophy that, although self-consciously *not* scientific, is shaped by theoretical goals and methods of reasoning that closely resemble those of the sciences. I'll be calling it "T-philosophy", to suggest "traditional" and "theoretical".

In line with his rejection of this approach, Wittgenstein's treatments of particular topics *within* the subject (including sensation, meaning, knowledge, logic, mathematics, religion, and art) don't seek to *answer* the usual T-theoretical questions—questions about the deep metaphysics and epistemology of these matters—but aim instead to show that they are *bad* questions and to take them off the table. This is to be done by showing that each one rests on its own misguided, topic-specific presuppositions, and showing moreover that these errors are fostered by the general ideology of T-philosophy.

The present account of Wittgenstein's perspective will proceed with discussions of: (i) the goals and methods that characterize T-philosophy; (ii) what I take to be a formidable Wittgensteinian argument against that practice; (iii) the sort of treatment of

[1] By Wittgenstein's "mature" ideas I mean the views articulated in his *Philosophical Investigations* (Oxford: Blackwell, 1953). Although strong anti-theoretical sentiments were already present in his *Tractatus Logico-Philosophicus* (London: Kegan Paul, 1922), they were given a cogent formulation only in the later work. See footnote 2 below.

particular philosophical problems that is called for by this argument; (iv) the issue of whether Wittgenstein's overall position is self-undermining—an anti-theoretical theory; (v) the question of whether it rests on an objectionable prioritization of *language* over *reality*, that is, an objectionable "linguistic turn"; and (vi) the difference between Wittgenstein's ideas and the Oxonian "ordinary language philosophy" of Austin, Ryle, Strawson, and others.

I should disclose from the outset that I am an admirer of Wittgenstein's (albeit a critical one) and think that his radical views should be taken seriously. To that end my principal aim here is *not* to offer an *interpretation* of his notoriously cryptic pronouncements but a *sympathetic development* of them. I'll be sketching and attempting to support a metaphilosophical position that I take to be plausible and important—one that's close to and suggested by what he explicitly says.

2. Goals of T-Philosophy

Work in this genre is oriented towards obtaining non-scientific theoretical knowledge about such matters as truth, necessity, justice, consciousness, reason, freedom, beauty, and existence. Some paradigm examples are Field's mathematical fictionalism, Davidson's view of meaning, Lewis' plurality of worlds, McTaggart's disproof of time, Tarski's definition of truth, Kant's ethics, and the integrated theory of metaphysics, representation, logical form, and metaphilosophy that Wittgenstein elaborated in his *Tractatus*.[2]

What are wanted are not mere collections of claims, but genuine *theories*. As in science, they qualify as such by purporting to be deep, explanatorily potent, true but not *obviously* true, and potentially revisionary with respect to our naïve opinions.

However, in contrast with science, T-philosophical theories aren't to be justified on the basis of *observation*. Conjectures of the sort ventured by philosophers and scientists working on naturalized epistemology, moral psychology, or empirical semantics, or the foundations of quantum mechanics, are excluded. But note that no disrespect towards investigations like these, nor any view to the effect that they "aren't *real* philosophy", need be attributed either to T-philosophers themselves or to their Wittgensteinian opponents.

Why those particular topics? Why truth, justice, etc., and not plastic, or football, or galaxies?—Partly, because the former phenomena strike us as puzzlingly *non-naturalistic* – as peculiarly hard to place within the vast network of objects and properties bearing spatial, temporal, causal, and explanatory relations to one another, and containing observable facts, the particles, fields, string, etc. for which those facts provide evidence, and all the 'bigger' things that are built out of such entities. But also because, insofar as a phenomenon isn't part of the naturalistic world, it won't manifest itself *observationally*, so it's theorization will have to be constrained by data of a different kind.

[2] This theory eventually became Wittgenstein's own primary example of the sort of philosophy he was against. In the Preface of his *Investigations* he describes the earlier book as pervaded with "grave mistakes". And one of them was the work's fundamental inconsistency—its anti-theoretical conclusion was based on theoretical premises! (For further discussion see chapter 3, sections 6 and 7, of my *Wittgenstein's Metaphilosophy*, Oxford University Press, 2012).

3. Methodology of T-Philosophy

The epistemological starting points for T-philosophical theorizing are our so-called intuitive judgments—for example, that planets must be located in space and time; that believing in ghosts is irrational unless supported by evidence; and that pigs can fly if and only if it's *true* that they can. Such facts play the role that in science is played by the observed data. However, there is no need to regard our beliefs about them as etiologically akin to observational beliefs—as the product of some sort of bizarre quasi-perceptual faculty. They may simply be equated with underived (i.e. epistemologically fundamental) a priori convictions.[3]

The next step—as in science—is extrapolation. Our intuitions about certain matters often appear to conform with simple and highly general principles—sometimes to such a striking extent that the correctness of those principles can seem undeniable—For example, that *any* real object is positioned in space and time; that *no* belief can be rational without evidence in its favour; and that *every* instance of "p" entails and is entailed by the corresponding instance of "It's true that p."

But such initial generalizations almost always run into trouble. Often what happens is that apparent counterexamples eventually spring to mind—conflicts with further compelling intuitions of ours: for example, that the number 3, although real, isn't the sort of thing that can be located in space and time, and that despite the absence of supporting argument or of any other form of evidence we are nonetheless justified in believing that if dogs bark then dogs bark. Or sometimes the problem is that our intuitively supported generalizations clash with one another: for example, the general principle of truth just-mentioned conflicts with the principles of classical logic (in that, combined together, they engender the liar paradoxes).

Within T-philosophy the responses to such tensions have taken a variety of forms. Most prominent amongst them are instances of: (i) *revisionism*, according to which certain pre-theoretical convictions about the phenomenon at issue are rejected (e.g. it's supposed that numbers must, after all, be material things); (ii) *skepticism*, the more extreme reaction of doubting or denying the very existence of the phenomenon (e.g. it's supposed that responsible, free choice is impossible); (iii) *mysterionism*, where the phenomenon (e.g. consciousness) is regarded as both perfectly real yet essentially paradoxical; and (iv) *conservative systematization*, according to which our initial generalization about the phenomenon

[3] My categorization of the data for T-philosophy as *a priori* is intended merely to reiterate the point that they are *not sensory*.—T-philosophical theses aren't to be justified on the basis of perceptual evidence. So it's worth emphasizing that even if the concepts of *a priori* and *a posteriori* should be abandoned (as Timothy Williamson suggests in his "How Deep is the Distinction between A Priori and A Posteriori Knowledge?" in Albert Casullo and Joshua C. Thurow (eds.), *The A Priori in Philosophy*. Oxford University Press, 2013, 291–312)—and even if they should not be used, as I have, to contrast perceptual beliefs (e.g. that this is red) with the sort of basic belief appealed to in T-philosophy (e.g. that only if Mars *is* red can I *know* that it is)—the prospects for an important contrast here remain highly plausible. Therefore the legitimacy of sharply separating T-philosophy from empirical science is not jeopardized by criticisms of the a priori/a posteriori distinction.

is recognized as overly simplistic, as falsified by some of the relevant intuitions, and it's acknowledged that what we need is a better theory—probably more complex—that will accommodate *all* the relevant data (e.g. utilitarianism must be modified to accommodate the "trolley problem").

4. Wittgenstein's Rejection of T-Philosophy

Wittgenstein's most sustained and developed account of his opposition to what I'm calling T-philosophy occurs in paragraphs 89 to 133 of the *Philosophical Investigations*. His main points, in a nutshell are:

1. There is a distinctive form of philosophical puzzlement, which has been taken to provide a distinctive motive for philosophical theorizing. The initial stimulus is not a straightforward matter of *curiosity*, as in science—that is, an awareness of ignorance and the desire to eliminate it. It's rather the conceptual tension of *paradox*, which engenders a sense of the phenomenon under scrutiny—e.g. free will, or consciousness, or truth, or meaning, or knowledge—as peculiarly baffling.
2. This sort of bewilderment arises (as we've just seen) because, on the one hand, generalizations occur to us whose combination of beautiful simplicity and capacity to capture our intuitive data makes them seem undeniable; but, on the other hand, these turn out to conflict either with one another or with further intuitive facts. We are left with the feeling that "This *must* be right, but *can't* be right."
3. Our sole aim should be to understand and dispel such characteristicall philosophical puzzlements. And to that end we must come to see that it's always the *theory* that's to blame.

As Wittgenstein puts it:

> Our investigation is . . . a grammatical one. Such an investigation sheds light on our problems by clearing misunderstanding away. Misunderstandings concerning the use of words, caused, among other things, by certain analogies between the forms of expression in different regions of language.
>
> (PI 90)

He is assuming (i) that what I've been calling our "intuitive judgements" or "basic a priori convictions" are just internalizations of "what we are inclined to say independently of perceptual evidence"; (i) that the T-philosophers' demand for simplicity leads them to unreasonably overstretch the "analogies between different regions of language", for example between our ways of using numerals and our ways of using planet names, between the uses of normative and non-normative terms, and between observation reports (such as "That is red") and first person sensation reports (such as "I have a pain in my left knee"); and (iii) that such internalized inclinations, explicitly articulated and generalized, become the tempting but irrational theories of T-philosophy.

In a similar vein:

> It was true to say that our considerations could not be scientific ones. It was not of any possible interest to us to find out empirically "that, contrary to our preconceived ideas, it is possible to think such-and-such"—whatever that may mean. . . . And we may not advance any sort of theory. There must not be anything hypothetical in our considerations. We must do away with all explanation, and description alone must take its place. And this description gets its light, that is to say its purpose, from the philosophical problems. These are, of course, not empirical problems; they are solved, rather, by looking into the workings of our language, and in such a way as to make us recognize those workings: in despite of an urge to misunderstand them. The problems are solved, not by giving new information, but by arranging what we have already known. Philosophy is a battle against the bewitchment of our intelligence by means of language.
>
> (PI 109)

As suggestive as I think these remarks are, they don't yet amount to a rationally convincing line of thought. Wittgenstein *claims* that a priori philosophical theorization should be abandoned, but doesn't adequately justify this claim: he doesn't show that such theories will invariably be untrue, or that they are bound to be irrational to believe, or even that it's unreasonable to expect, at the outset of any such investigation, that interesting knowledge will be obtained.

I will attempt in what follows to supply some supporting arguments. I don't think they can be found in Wittgenstein's writings, but I believe they provide a natural and necessary extension of what he does say.

5. The Case Against T-Philosophy

The bottom-line Wittgensteinian complaint is that T-philosophy's primary goal cannot be achieved: the theoretical knowledge to which it aspires isn't to be had. This conclusion rests on a combination of three broad considerations: (i) that our intuitive judgments defy *simple* superficial systematization; (ii) that the techniques enabling *scientists* to devise simple and plausible theories in the face of their equally messy data can't work in T-philosophy; and (iii) that although temporarily adequate, *complex* systematizations *can* be devised in T-philosophy, there will typically be a set of alternative proposals that are equally adequate, with only rare rational convergence on just one of them, and no objective value in arriving at the truth. Let me go over these points in more detail.

(I) Our intuitive judgments—our basic a priori convictions—are intractably messy. Not only do we discover this to be so, but it should be unsurprising that we do. For the linguo-conceptual practices that are explicitly articulated in such judgments evolved under a variety of shifting and often conflicting constraints (—including our practical and intellectual needs, our limited cognitive powers, and our physical and social environments). So we are almost always left with concepts that are too complex, and/or too vague, and/or too open ended, and/or too family-resemblance-like, and/or too paradox-prone, to be captured by simple definitions, or rules, or guiding theories.

(II) Of course, the data in science are no less dauntingly complex and no less resistant to neat superficial systematization—yet simple adequate theories are regularly found. And a non-scientific philosopher might well be encouraged by this observation to expect the same thing. But, on reflection, that must be regarded as a vain hope. An examination of the various routes to simplicity enabling the *scientist* to make satisfying discoveries shows that they won't be available to T-philosophers.

Consider, first, the common and fruitful scientific practice of explaining the characteristics of *whole* systems in terms of the properties of their *parts* and how those parts are arranged with respect to one another. For example, macroscopic phenomena are explained in terms of their microscopic components, and the features of a community are explained in terms of the properties of its individual members. Substantial gains in simplicity are often achieved in this way because a plethora of complex systems, each one tending to behave somewhat differently from the others, often turn out to be constructed by different arrangements of relatively few common parts—in which case we can devise a theory that explains the diverse behaviours of the many complexes in terms of relatively few assumptions about those parts and about the consequences of alternative arrangements of them.

What prevents this strategy of explanation from being deployable within T-philosophy is that it can apply only when the phenomena under investigation are complex systems whose characteristics issue causally from their internal components and from the spatial relations of those components to one another. So it's hardly surprising that such explanations of *non-naturalistic* reality never proposed. What division into 'parts' could conceivably be relevant to the search for a simple T-philosophical account of when actions are free, of what it is to have knowledge, of which principles govern a just society, etc.?

Granted, there do exist proposals in T-philosophy that are *somewhat* analogous to the explanation of wholes in terms of their parts. One, of course, is *conceptual analysis*: the a priori deconstruction of a phenomenon—for example, knowledge, or free action, or truth, or goodness. But notice that such proposals are not typically offered as attempts to give a simple adequate account in the face of more superficial generalizations that aren't able provide such a thing. On the contrary, it's the suggested philosophical analyses *themselves* that almost always fail to do justice to the complexity of the relevant intuitive facts. These analyses don't *solve* the problem of our inability to capture our messy data—they *are* the problem.[4]

Prominent in the philosophy of language is another sort of T-philosophical theory that's inspired by part–whole explanation. The idea is to explain the properties of sentences—especially their truth conditions and their meanings—in terms of properties attributed to

[4] An arguably successful case of a priori reductive theorizing is Frege's proposal (in his *The Foundations of Arithmetic*) to explain arithmetic in terms of logic and set theory. The positive integer, n, is identified with the set of all n-membered sets. But here also there is no significant simplification. What are purportedly explained are the Peano axioms of arithmetic—which are already simple. Moreover, we have entirely lost the idea of wholes explained by their parts. Indeed, and partly for that reason, it's far from clear that sets are in any sense more fundamental than the numbers they are supposed to be constituting.

their constituent words (e.g. their referents or their conceptual roles) and in terms of the logico-syntactic structures in which the words are embedded.

Here the analogy with science is more solid—we really are offered a sort of part–whole explanation. However, just as in the case of conceptual analysis, a crucial disanalogy is that these proposals are not geared towards the unification of phenomena that resist more superficial systematization. On the contrary, the facts to be explained are rather easy to systematize—we already have the relatively uncontroversial schemata,

"p" is true if and only if p

and

"p" expresses the proposition *that p*

The availability of these accounts casts considerable doubt both on the need for the compositional theory and on the prospects for devising one that will be adequate.

Besides part–whole explanations, a second strategy that is commonly deployed within science to help systematize messy data is to see these facts as the product of simple *tendencies* together with distorting factors—or in other words, the product of *ceteris paribus* laws plus circumstances that lead to divergences from them. For example, we get a simplifying handle on the messy facts concerning headaches and aspirin by recognizing that *aspirin tends to remove headaches*. Similarly, the rather erratic weather in New York might be theorized in terms of the law that *its November days tend to be rainier than its July days*.

But this approach—like the previous one—is too *causal* to be applicable in non-naturalistic contexts. No doubt there are simple T-philosophical generalizations that hold very often but not always. That's precisely what the phenomenon of "complex intuitive data" amounts to. However, talk of "laws" or "tendencies" is surely out of place. If someone is persuaded by the existence of numbers and values that the world is not entirely naturalistic, she couldn't plausibly retreat to the claim that there's an other-things-being-equal *law* to the effect that reality is naturalistic. Similarly, supposing that *normative* beliefs provide a counterexample to the general view that beliefs can't be intrinsically motivating, it wouldn't remain plausible that there's a mere *tendency* for beliefs to lack that psychological power.

In addition to part–whole explanations and invocations of tendencies, there's a third route to simplicity in science: namely, *idealization for explanatory purposes*. An underlying model, based on explicit simplifying assumptions that are known to be false, is used to explain why a relatively superficial lawlike generalization is roughly right. For example, the supposition that a gas is composed of *point* masses was used, in conjunction with statistical mechanics, to explain the approximate correctness of Boyle's Law. Similarly, the supposition that the planets exert no gravitational forces on one another was used to explain the rough truth of Kepler's Laws.

In such cases, the scientists usually have a fairly good idea of how the deliberately ignored complications could *in principle* be brought into the picture to engender a more accurate explanation of the phenomena. But the *details* about these factors may be hard to know, and may vary considerably—even randomly—from one instance to

another.[5] Or the calculation of their effects may be mathematically intractable. So the scientist is content with her demonstration that if things *were* "ideal" in such-and-such a specified way, then such-and-such laws *would* hold and be explained—inferring from this that the *approximate* correctness of the idealization really does explain why those laws are *approximately* correct, and comforted by the knowledge that the full story is in practice unobtainable only because of its complexity and not because anything very important remains to be understood.

But it is doubtful that any such thing is ever done, or can be expected, in T-philosophy. For, as we have seen, genuine explanatory depth in science is achieved by accounts of wholes in terms of their parts, within the context of a fundamental a posteriori and naturalistic theory of space, time, force, and motion. And no such explanations are conceivable within T-philosophy.

However, it will reasonably be protested that we surely *do* come across idealizations in philosophy! Consider, for example, the supposition of our brains being infinitely large, enabling instantaneous computations of all the theorems of the predicate calculus, so that rationality would require us to be absolutely certain of their truth.[6]

Well, we must concede that this idealization may have its virtues for certain broadly philosophical purposes. Perhaps it could earn its keep in *naturalized* epistemology, within an empirical study of how we reason. And even more promising is its role in the Bayesian probabilistic model of rational degrees of belief—a model whose nice blend of simplicity and approximate truth helps to remove confusions and paradoxes in the philosophy of science (e.g. the raven paradox). This, it seems to me, would be a worthwhile attempt at Wittgensteinian therapy.[7]

But, in contrast to both of these aspirations, the T-philosopher aims to specify the fundamental normative principles that characterize *actual* human rationality. And it's hard to imagine how the above, idealized account is going to lead him in the direction of that goal—towards a theory in which the idealization has been removed. Granted, it might explain the approximate truth of the central Bayesian norm ("One's degrees of belief should conform to the probability calculus"); and in this respect it does parallel the explanatory rationale for scientific idealizations. But the crucial difference is that the scientist very often has a rough understanding of how, at least in principle, to dispense with her idealization and incorporate the non-ideal factors into a completely accurate story—whereas the T-philosopher hasn't the foggiest idea of how to do that.[8]

[5] Idealizations often take the form of assuming that the systems in question are *isolated*—i.e. that no external factors are impinging on them. And a good part of the reason for this is that, in reality, many such factors vary *randomly*.

[6] See Timothy Williamson's review of my *Wittgenstein's Metaphilosophy* in the *European Journal of Philosophy* 21 (S2):e7–e10 (2013).

[7] This is argued in "Wittgensteinian Bayesianism," *Midwest Studies in Philosophy*, Volume 18, edited by P. French, T. Uehling, and H. Wettstein, University of Notre Dame Press, 1993, pp. 62–77 (reprinted with revisions as chapter 5 of my *From a Deflationary Point of View*, Oxford University Press, 2004).

[8] No doubt a formal articulation of the idealized model in terms of probability theory will engender some challenging logical and mathematical problems, and philosophers frustrated with the vagaries and inconclusiveness of real T-philosophy may derive satisfaction from solving them, from getting

Another multiply interpretable idealization is *formal logic*, as developed initially by Frege and Russell, and exploited in Wittgenstein's *Tractatus*. The idea was not the construction, for its own sake, of a new language, together with stipulated rules of deduction for its sentences. Those so-called "logical formulae" were supposed to bear an intimate relation to ordinary language; they were supposed to articulate the "conceptual contents" or "pure thoughts" expressed by ordinary sentences—leaving out the distracting complexities of linguistic features that are present for merely pragmatic reasons (e.g. merely for emphasis, or for the expression of some conative attitude toward the content that's being simultaneously asserted). Thus logic was seen as an idealization of ordinary language.

But what functions might this idealization have, and which theoretical assumptions would they presuppose? As in the Bayesian case, there are three possibilities. One is to treat it as a *scientific* idealization, along the lines of Chomsky's LF hypothesis in empirical linguistics. Another is to regard logic as combining *similarity to natural language* with *clarity and simplicity* in just the right way to facilitate the detection of fallacies in our natural-language reasoning (especially in philosophy!). In the spirit of the later Wittgenstein, it would be given a merely instrumental and therapeutic function. And finally it could be a T-philosophical idealization—the proposed first step towards accurate and complete a priori accounts of the natures of truth, meaning, and sound inference. But think of the battery of complications that will need to be eventually accommodated: vagueness; context-dependence; pragmatics; the differences in meaning between the formal symbols ("–", "→", …) and the corresponding ordinary words ("not", "if", …); the many ordinary constructions that resist logical formalization (e.g. belief attributions, generics, and laws of nature); … and so on. Only the part–whole explanations of science—yielding reductions of linguistics to neurology to biology, etc., and bottoming out in a fundamental physics—can offer any prospect of explaining such phenomena (or of determining that the logical idealization is not after all of any value). Thus these matters call for empirical investigation and causal theorization. An a priori philosophical idealization isn't going to yield anything beyond vacuous stipulations.[9]

A fourth strategy—the only other I can think of—that is sometimes deployed by scientists for the sake of theoretical simplicity is to maintain that some of the data should be thrown away: that some of the reported observations must be *mistaken*. And roughly this sort of move is indeed made within T-philosophy. It *is* often argued there that a certain theoretical generalization is so wonderfully simple and so nearly consistent with all the intuitive data that the apparently falsifying information must be spurious. In fact such thinking lies behind all but one of the prominent forms of T-philosophical theory (listed in section 3.1)—namely, revisionist theories, skeptical theories, and mysterionist theories. For example, it's supposed that considerations of data-fit and simplicity compel us to accept that:

definite answers to formal questions. But it's not at all obvious that these answers will help us towards the above-mentioned goal. And what would be the point of rigor without relevance? (This footnote is taken from my "Reply to Timothy Williamson's Review of *Wittgenstein's Metaphilosophy*," *European Journal of Philosophy* 21(S3), e18–e26, 2013).

[9] These points about the relationship between logic and language are suggested by paragraphs 81, 89, 97, and 107 of the *Philosophical Investigations*, where Wittgenstein condemns the theory that he had advanced in his *Tractatus*. For example: 'The more narrowly we examine actual language, the sharper becomes the conflict between it and our requirement. (For the crystalline purity of logic was, of course, not a result of investigation: it was a requirement). The conflict becomes intolerable; the requirement is now in danger of becoming empty. …' (PI 107)

> The rightness of an act depends solely on the overall amount of pleasure (and absence of pain) that it should be expected to bring about.—So we must abandon our intuition that it would be wrong to kill someone in order to use his organs to save a large number of others. (Revisionism)
>
> No belief is intrinsically motivating.—And so much the worse for our sense that whenever we are fully convinced we *ought* to do a given thing we have some inclination to do it. (Revisionism)
>
> All objects and properties are *naturalistic*, i.e. elements of the causal explanatory spatiotemporal nexus.—Therefore, since numbers obviously wouldn't be things of that sort, they couldn't exist. (Skepticism). They clearly do exist; so it must *somehow* be possible for certain things to be non-natural; but we'll never be able to understand how that could be so. (Mysterionism)

These proposals involve commitments to generalizations that can be sustained only by supposing that their simplicity legitimizes an overriding of intuitive judgments. And T-philosophy is pervaded with this sort of thing.

But on reflection we can see that *its* practice of data dismissal is only superficially similar to what's done in science. It isn't as rationally disciplined. In science, any discounting of observations must be *independently* justified. It must be shown, for example, that their circumstances make them unreliable, or that they are widely disputed, or that they presuppose a controversial theory. No doubt there are cases in T-philosophy too in which what would normally be theory-constraining judgments are legitimately discounted for one of those reasons. But, as illustrated by the examples that we have discussed, it's not at all abnormal for T-philosophers to act as though simplicity can trump awkward data without further ado.

Can this contrast with science perhaps be explained and rationalized by the comparative *weakness* of a priori intuitive beliefs? Are they intrinsically less compelling than observational beliefs, so that it's neither impossible nor unreasonable to ignore them for the sake of massive gains in simplicity? I don't think so. For the present issue concerns only those intuitive judgments that we nearly all regard as *certainly* correct: for example, that Julius Caesar was not a number, that $\langle 1 + 2 = 3 \rangle$ is true $\leftrightarrow 1 + 2 = 3$, that it's wrong to grab someone off the street in order to harvest his organs, etc. To think that such data should be suppressed for the sake of theoretical simplicity is an irrational perversion of the scientific practice.

It must be conceded that, even in science, a theory is sometimes accepted despite its inconsistency with *unimpeachable* data. For it is expected that one day the anomalies will be explained away. So what's wrong with doing the same thing in T-philosophy? The answer is that when the scientific datum at issue really is incontrovertible—if, for example, all observers agree that the needle points to "3.1" instead of to the predicted "4.2"—then the expected "explaining away" would be a matter of finding something wrong with one of the many *auxiliary assumptions* that are almost invariably needed to derive a prediction from a theory. But in T-philosophy the theories proposed are inferentially relatively close to the data, and so the prospects for being able to hang on to both the generalization and the recalcitrant facts are very limited.[10]

(iii) In light of the difficulties we have just surveyed of reconciling a simple theory in T-philosophy with the messy relevant data, the honest conclusion—accepted by many

[10] Subsection (ii) is a substantial expansion of the explanations I've offered before of why none of the legitimate techniques for arriving at simple theories in science are available in T-philosophy. My earlier accounts of this were in chapter 2, section 4, of Wittgenstein's Metaphilosophy and in my above cited "Reply to Timothy Williamson's Review …".

of its practitioners—is that an adequate theory might well have to be disappointingly complex. But now the question arises as to whether a highly complex theory could be made sufficiently plausible to count as knowledge. The problem here—a familiar observation in the philosophy of science—is that the lower the level of simplicity required, the greater the number of theories that will qualify. So although one might succeed in contriving a *convoluted* systemization of the intuitive date in some area, there are probably going to be alternatives to it; our epistemic norms won't enable us to decide between them; so objective knowledge of which of them is correct won't be achievable. Consider:

- The contest between Aristotelean, Kantian, utilitarian and other theories of morality
- The many competing accounts of knowledge
- The explosion of so-called "solutions" to the liar paradoxes

These examples illustrate the normal situation in T-philosophy—a continual profusion of competing accounts, each one elaborated in a variety of different directions. What we don't see is an eventual convergence on some theory whose unique blend of adequacy and simplicity make it a plausible candidate for knowledge.

Nonetheless, we might object, it's surely not irrational, if one happens to derive great pleasure from doing so, to *try* to devise a simple theory in some domain and, after failing to come up with one, to confine oneself to exploring the space of alternative complex theories. And that's obviously right. It's merely an instance of the prima facie reasonability of attempting to satisfy one's desires.

But what remains irrational is: (i) to expect the initial search for a simple theory to be successful; (ii) to expect that the subsequent exploration of alternatives would have any *objective* value; (iii) to hide from oneself that one is probably motivated by an unreasonably high confidence that one's favorite theory *will* eventually be established; and (iv) to think that if some particular highly contrived account *were* somehow to emerge as better supported than the others, then, despite the convolutions, its truth would still be worth knowing.[11]

Besides the various detailed reasons just elaborated for doubting that T-philosophy can be successful, there is another simple but weighty consideration. Namely, that the 'burden of proof' here is on those who assume (explicitly or implicitly) that T-philosophy *can* deliver the sort of knowledge to which it aspires—yet there is no rational basis for that assumption. Pretty obviously one shouldn't be moved, merely by the example of the natural sciences, to suppose that worthwhile theorization of *non*-natural phenomena can equally well be expected. For, not only are these two broad domains of investigation radically different from one another in relevant respects, but their contrasting records to date—impressive success in one and impressive failure in the other—can hardly be ignored. So what then might be said in favor of T-philosophy? Absent a decent answer, this endeavor should not be regarded as objectively valuable. Thus a good case against it doesn't require a direct argument to the effect that the understanding it seeks to *cannot* be achieved. It suffices that we lack the slightest reason to think that it *can* be achieved.

[11] For discussion of the common-sense idea that not *all* true belief is valuable, see my "Belief-Truth Norms" in T. Chan (ed.) *The Aim of Belief,* Oxford University Press, 2013.

6. How to Treat Specific Topics *within* Philosophy

The above-sketched characterization and critique of T-philosophy suggests that the sources of a philosophical problem together with the strategies for its resolution may usefully be organized by means of the following sequential schema:

> Our *scientistic demand* for simple a priori accounts combines with a recognition of certain *linguistic analogies* between how X-terms and words in other domains are used, resulting in a compelling *generalization*. But this turns out to clash with certain *idiosyncrasies* in our use of those terms. That *paradoxical tension* is addressed within T-philosophy by means of a variety of alternative strategies, most prominently: *revisionary*, *skeptical*, *mysterionist*, or *conservative*. But such accounts are predictably unsatisfactory. What's needed instead is a *therapeutic dissolution* of the paradox in which the defects of the various theoretical proposals are exposed, and resolution is found instead in an abandonment of the irrational scientistic overgeneralization that caused the problem in the first place.[12]

The idea that this schema (with some minor variations) illuminates *all* T-philosophical problems is perfectly consistent with Wittgenstein's methodological pluralism (e.g. his remark in PI 133 that "There is not *a* philosophical method, though there are indeed methods, like different therapies"). Nor can it be objected that our treatments of specific problems would become mindlessly mechanical. For a belief that the abstract framework can helpfully be applied to a given case will be no substitute for the hard, detailed work of filling in the schematic categories with substantial claims (which will vary enormously from problem to problem) about exactly what the overstretched analogies are in that case, which are the paradoxes to which they give rise, what specifically is wrong with each of the alternative a priori theories commonly proposed to resolve the problem, and which particular techniques (including reminders about ordinary usage) are best for dissolving our puzzlement.

For concreteness, let me very sketchily indicate how the schema might be filled out in the case of the traditional question, "What is truth?". This question provokes perplexity because, on the one hand, it demands an answer of the form, "Truth is such–and-such," but on the other hand, despite hundreds of years of looking, no acceptable answer of that kind has ever been found. We've tried truth as "correspondence with the facts," as "provability," as "practical utility," and as "stable consensus"; but all turned out to be defective in one way

[12] Some will insist that the best alternative to T-philosophy, and what we should really do if T-philosophy is indeed hopelessly misguided, is *not* Wittgensteinian therapy but *naturalistic* philosophy á la Quine. However, it's worth emphasizing that if, as Wittgenstein suggests, traditional philosophy has been largely driven by a distinctive form of puzzlement—the bewilderment of being stuck in paradox—and if, as he claims, such conceptual tensions are the product of our irrational presuppositions rather than scientific ignorance, then the way forward must include the painstaking therapeutic self-examinations that are illustrated by his treatments of meaning and sensation in the *Investigations*, and his treatment of knowledge in *On Certainty* (G. E. M. Anscombe and G. H. von Wright (eds.), Oxford: Blackwell, 1969). Science may well be able to answer important questions about the phenomena at issue, but it isn't designed to resolve the self-inflicted problems (or pseudo-problems) most characteristic of our subject.

or another—either circular or subject to counterexamples. Reactions to this impasse have included a variety of theoretical proposals. Some philosophers have been led to deny that there is such a thing as absolute truth (*skepticism*). Some have maintained (insisting on one of the above definitions) that although truth exists, it lacks certain features that are ordinarily attributed to it (*revisionism*)—for example, that the truth may sometimes be impossible to discover. Some have inferred that truth is intrinsically paradoxical and essentially incomprehensible (*mysterionism*). And others persist in the attempt to devise a definition that will fit all the intuitive data (*conservative systematization*).

But from a Wittgensteinian perspective each of the first three of these strategies rides roughshod over our fundamental convictions about truth, and the fourth is highly unlikely to succeed. Instead we should begin, he thinks, by recognizing that our various concepts play very different roles in our cognitive economy and (correspondingly) are governed by defining principles of very different kinds. Therefore, it was always a mistake to extrapolate from the fact that empirical concepts, such as RED, or MAGNETIC, or ALIVE, stand for properties with specifiable underlying natures to the presumption that the notion of TRUTH must stand for some such property as well. Wittgenstein's conceptual pluralism positions us to recognize that notion's idiosyncratic function, and to infer that neither the concept, TRUTH, nor the phenomenon of truth itself, will be reducible to anything more basic. More specifically, we can see that the concept's function in our thought and talk is merely to serve as a device of generalization.—It enables us to say such things as "Einstein's last words were true," and not be stuck with "If Einstein's last words were that $E=mc^2$, then $E=mc^2$; and if his last words were that nuclear weapons should be banned, then nuclear weapons should be banned; …, and so on"—which has the disadvantage of being infinitely long! Similarly we can use it to say "We should want our beliefs to be true" (instead of struggling with "We should want that if we believe that $E=mc^2$, then $E=mc^2$; and that if we believe ..., etc."). We can see, also, that this sort of utility depends upon nothing more than the fact that the attribution of truth to a statement is trivially equivalent to the statement itself—e.g. "It's true that $E=mc^2$" is equivalent to "$E=mc^2$.". Thus possession of the concept of TRUTH appears to consist in an appreciation of that triviality, rather than a mastery of any explicit definition. The traditional search for such an account (or for some other form of reductive analysis) was a wild-goose chase, a pseudo-problem. Truth emerges as exceptionally un-profound and as exceptionally un-mysterious.[13]

7. Is Wittgenstein's Metaphilosophy Incoherent?

A commonly voiced objection to Wittgenstein's anti-theoretical perspective is that it's itself a theory, hence self-undermining. But this charge doesn't stand up to scrutiny.

To see why not, notice first that not every claim can correctly be described as "theoretical" in the ordinary sense of that word (which is the sense that Wittgenstein is using). A *theory* properly so-called is something whose truth cannot be made obvious, but can be

[13] The preceding paragraphs on truth are extracted from my online article, "Was Wittgenstein Right?", which appeared in the *New York Times* philosophy blog, *The Stone,* on March 3, 2013. For more detail, see chapters 1–5 of my *Truth-Meaning-Reality*, Oxford University Press, 2010.

justified only by some form of conjectural inference.[14] It's clear then that not even every *controversial* claim is theoretical. For something may be potentially obvious, but nonetheless denied by those who are confused or not looking in the right direction. So the mere fact that T-philosophers reject Wittgenstein's critique does not make it a theory.

But even if it *is* a theory, that wouldn't be enough to convict him of hypocritically engaging in the very intellectual activity he's condemning. In order for such an accusation to stick, it would have to be shown that his theory is an example of *T-philosophy*. More specifically, it would have to be shown that his own critical observations themselves rest on just the sort of dismissals of data or a priori conjecturing that he is complaining about in T-philosophy. But none of the accusers have even attempted to make that case—and as far as I can see it can't be made. The considerations against T-philosophy that were lodged in section 5) rest entirely on a combination of plausible a posteriori generalizations (e.g. that our concepts unsurprisingly resist systematization) and uncontroversial norms of reason (e.g. that it's irrational to dismiss data solely on the grounds that they would falsify a simple theory).[15]

8. Does He Take an Objectionable "Linguistic Turn"?

Another familiar gripe about Wittgenstein's perspective is that he wrongly focuses on *language* rather than *reality*. And indeed he does repeatedly tell us that philosophical puzzlement derives from our being tricked by language—in particular, that we're tempted to overstretch analogies between the uses of words in different domains of discourse.[16] His idea is, for example, that our puzzlement about numbers derives from an exaggeration of the similarity in use between numerals (e.g. "3") and names of physical objects (e.g. "Mars"); and that our puzzlement about truth derives (as we have just seen) from an exaggeration of the similarity between our use of the word "true" and our use of empirical predicates such as "red" and "magnetic." However, one might complain that if there are any mistakes here they aren't really about language but about the world itself—that the analogies we overstretch are actually between the *objects* 3 and Mars, and between the *properties* truth and redness.

This objection strikes me as overstated. It is true that one shouldn't think our exaggerations concern language *rather than* reality. But nor do they concern reality *rather than*

[14] "if we too in these investigations are trying to understand the essence of language—its function, its structure—yet *this* is not what those questions [raised and addressed in the *Tractatus*] have in view. For they see in the essence, not something that lies open to view and that becomes surveyable by rearrangement, but something that lies *beneath* the surface. Something that lies within, which we see when we look *into* the thing, and which an analysis digs out" (PI 92).

[15] See chapter 2, section 8, of *Wittgenstein's Metaphilosophy* for a more detailed discussion of this issue. A related objection often made against Wittgenstein is that his famous identification of the *meaning* of a word with its *use* in the language is an instance of T-philosophy. My own response (elaborated in chapter 4, section 2d) is that, according to Wittgenstein, this "definition" of meaning (to use his term) really is just a *definition*: that it really is implicit in our fundamental use of the term "meaning"; implicit in our refusal to allow that someone means what we do by a word if her basic use of it is deeply different from ours; that it's no more of a speculative conjecture than is any other proposed definition (e.g. of "bachelor"); and so it isn't a theory of the sort he thinks must be avoided.

[16] See PI 90 and PI109, quoted in section 4.

language. For, quite clearly, we're doing both at the same time. Our *linguistic* over-generalizations go hand in hand with over-generalizations at the *material* level. After all, a person's accepting the sentence "p" is what constitutes her commitment to the *fact* that p. So analogies—whether real or exaggerated—in *use* (i.e. in accepted sentences containing them) between the words "f" and "g" are bound to coincide with analogies in the *supposed facts* involving the entities f and g themselves. A similarity in how one uses the terms "5" and "Mars" coincides with a similarity that one takes to exist between 5 and Mars.

And I think it would be uncharitable to accuse Wittgenstein of failing to see all this. By "our language" he means "our linguo-conceptual activity"; and his real point is that our problems derive from the occasional irrationality of this activity and are to be dissolved by critical reflections designed to identify and remove those mistakes. It doesn't matter whether we describe the activity as "acceptance of sentences" or as "material commitments."

Another *apparently* objectionable expression of "prioritization of language" is his tendency to maintain that philosophical theories are *linguistically* defective.[17] For he's often construed here as retaining his implausible early contention (in the *Tractatus*) that all such theories are *meaningless*. But, despite his own occasional potentially misleading ways of putting the matter, again I'd say that this is not his intent. As already suggested, the defect he has in mind isn't really *strict nonsensicality* but *irrationality*. T-philosophical theories *don't* usually have any effect at all on meanings. (For example, "An act is *good* if it brings about the greatest happiness of the greatest number," and "*Truth* has a specifiable underlying nature"). And even when they do have such effects, the results aren't *no* meanings but *new* meanings. Consider, for example, Lewis's "Any merely possible event is nonetheless *real*", Tarski's "It's $true_2$ that Goldbach's Conjecture is $true_1$", and Heidegger's "*The nothing noths.*" Since, for the later Wittgenstein, "meaning = use," and since in such cases we surely do have genuine cases of words being given new, technical uses, then he can hardly deny their meaningfulness. Yes, the meanings they have are *irrational* to deploy, given the irrationality of the theory-driven uses underlying them—but they are meanings nonetheless.

9. "Ordinary Language Philosophy"

Some of the ideas that Wittgenstein was developing and disseminating at Cambridge University during the 1930s and 40s gradually made their way across to Oxford and had an obvious (though often unacknowledged) influence on some of the leading philosophers there.[18] Ryle, Austin, Strawson, and others, took to heart his emphasis on the vital

[17] "Where does our investigation get its importance from, since it seems only to destroy everything interesting, that is, all that is great and important? (As it were all the buildings, leaving behind only bits of stone and rubble). What we are destroying is nothing but houses of cards and we are clearing up the ground of language on which they stand" (PI 118).

"The results of philosophy are the uncovering of one or another piece of plain nonsense and of bumps that the understanding has got by running its head up against the limits of language. These bumps make us see the value of the discovery" (PI 119).

[18] See Lynd Forguson's "Oxford and the 'Epidemic' of Ordinary Language Philosophy," *The Monist* 84:3, July 2001, 325–45.

importance to philosophy of paying scrupulous attention to the ordinary usage of words—especially to the usage of those words denoting phenomena of philosophical concern, e.g. "proposition," "pain," "good," "intend," "know," "free," and so on. For they were persuaded that the prevalence of poor philosophical theories—either blatantly false if construed literally, or pointlessly idealized if not—resulted from the lack of such attention.

But it would be misleading to lump Wittgenstein together with these others as members of the same intellectual movement, "ordinary language philosophy", since most substantially diverge from him on the question of what's to be done in light of the agreed points. For most of the Oxford philosophers, the upshot was that good philosophical theories could and should nonetheless be devised and established, but would have to be considerably more complicated than had previously been recognized. Thus their goal can be assimilated to that of *conservative systematization*, which we criticized from a Wittgensteinian perspective at the end of section 5.1.

For him, the massive complexity of language—hence the extreme messiness of our intuitive data—shows that no interesting theoretical results can be expected. He saw no point in attempting to construct such theories—not merely because their inevitable convolutions and ad hoc features would generally make them impossible to establish, but also because no such theory—no contrived and ramshackle collection of principles—would be objectively worth knowing the truth of. Therefore his sole aim was to resolve the apparently inescapable conceptual tensions that typify T-philosophy. He saw that these arise from the mismatch between, on the one hand, the messiness of the phenomena and, on the other hand, the almost irresistible allure of simplicity and its upshot in irrational over-generalization. So he advocated a form of therapy as a result of which these temptations would be acknowledged and held in check. And that would be the end of the matter:

> Our craving for generality has [as one] source ... our preoccupation with the method of science. I mean the method of reducing the explanation of natural phenomena to the smallest possible number of primitive natural laws; and, in mathematics, of unifying the treatment of different topics by using a generalization. Philosophers constantly see the method of science before their eyes, and are irresistibly tempted to ask and answer in the way science does. This tendency is the real source of metaphysics, and leads the philosopher into complete darkness. [19]

Thus he would say that the Oxford "ordinary language philosophers" had grasped *part* of his point, but not the most important part.[20, 21]

[19] Wittgenstein's *The Blue and Brown Books* Oxford: Blackwell, 1958, p. 18.

[20] Perhaps he had some of these people in mind, amongst others, when he wrote in the 1945 Preface to his *Investigations*: "Up to a short time ago I had really given up the idea of publishing my work in my lifetime. It used indeed to be revived from time to time because I was obliged to learn that my results (which I had communicated in lectures, typescripts and discussions), variously misunderstood, more or less mangled or watered down, were in circulation. This stung my vanity and I had trouble quieting it."

[21] This paper is based on material I've presented on various occasions: a workshop on "Expressivism, Pragmatism, and Realism" at the University of Sydney (June 2011), the meeting "In Memory of Richard Rorty" in Buenos Aires (May 2013), a conference on "The Nature of Philosophy" at the University of Dublin (June 1913), the Inter-American Congress of Analytic Philosophy in Salvador, Brazil (October 2013), Cambridge University's Moral Sciences Club (November 2013), and a workshop on "Wittgenstein's Contemporary Relevance" at the American University of Beirut (May 2014). I would like to thank the audience members on these occasions for their tough questions. In addition I'm most grateful for the generous and invaluable feedback I've received in discussions with Saleh Agha, Paul Boghossian, Tim Crane, Penn Lawrence, Huw Price, Stephen Schiffer, Meredith Williams, Michael Williams, and Timothy Williamson.

References

Forguson, L. "Oxford and the 'Epidemic' of Ordinary Language Philosophy," *The Monist* 84:3, July 2001, 325–45.

Frege, G. *The Foundations of Arithmetic*, J. L. Austin (tr.), Evanston, Ill., Northwestern University Press, 1980.

Horwich, P. "Wittgensteinian Bayesianism", Ch. 5 of his *From a Deflationary Point of View*, Oxford, Oxford University Press, 2004, 105–27.

Horwich, P. *Truth-Meaning-Reality*, Oxford, Oxford University Press, 2010.

Horwich, P. *Wittgenstein's Metaphilosophy*, Oxford, Oxford University Press, 2012.

Horwich, P. "Reply to Timothy Williamson's Review of *Wittgenstein's Metaphilosophy*", *European Journal of Philosophy* 21(S3), 2013, e18-e26.

Horwich, P. "Was Wittgenstein Right?", *The Stone (New York Times* online philosophy blog) on March 3, 2013.

Horwich, P. "Belief–Truth Norms", T. Chan (ed.) *The Aim of Belief*, Oxford, Oxford University Press, 2013, 17–31.

Williamson, T. "*Wittgenstein's Metaphilosophy* – By Paul Horwich", *European Journal of Philosophy* 21(S2), 2013. e7-e10.

Williamson, T. "How Deep is the Distinction between A Priori and A Posteriori Knowledge?", A. Casullo and J. C. Thurow (eds.), *The A Priori in Philosophy*, Oxford, Oxford University Press, 2013.

Wittgenstein, L. *Tractatus Logico-Philosophicus*, London, Kegan Paul, 1922.

Wittgenstein, L. *Philosophical Investigations*, Oxford, Blackwell, 1953.

Wittgenstein, L. *On Certainty*, G. E. M. Anscombe and G. H. von Wright (eds.), Oxford, Blackwell, 1969.

Wittgenstein, L. *The Blue and Brown Books*, Oxford, Blackwell, 1958.

CHAPTER 8

PHILOSOPHICAL NATURALISM

HILARY KORNBLITH

1. Introduction: What is Naturalism?

One might have thought that philosophers who regard themselves as naturalists would be in agreement about proper method in philosophy. There is, however, no such agreement to be found. What one finds, instead, is a great diversity of opinion about proper method, as well as a great diversity in the actual practice of naturalistically minded philosophers. This chapter surveys the range of naturalistic approaches to philosophical methodology.

Naturalism is often defined as a methodological doctrine. Those who favor this view of naturalism often quote with approval W. V. O. Quine's remark that philosophy should be seen as "continuous with science" (Quine, 1969, 126). Unfortunately, this quotation is not fully clear on what a proper naturalistic method in philosophy might be. Some kind of deference to science is obviously intended, but exactly how deep that deference might go is entirely unclear. To say that philosophy should be "continuous with science" does not, by itself, commit one to any particular methodology. There are stronger and weaker readings of the continuity claim, ranging from the view that philosophical method is no different at all from the methodology of science, and, indeed, that philosophy should be absorbed into or replaced by one or more of the sciences, to views which allow for substantial departures in method between philosophy and science as long as those differences are matters of degree rather than matters of kind.

Those who see naturalism as primarily a metaphysical doctrine often quote Wilfrid Sellars with approval: "science is the measure of all things, of what is that it is, and of what is not that it is not" (Sellars, 1963, 173). It is even more obvious here that a wide range of methodologies is compatible with such a metaphysical commitment.[1] For one thing, there

[1] It is especially worth pointing out that the quotation from Sellars is typically removed from some very important features of its context, as I did—following tradition. A fuller quotation from Sellars reads as follows: "in the dimension of describing and explaining the world, science is the measure of all things."

is disagreement among naturalists about which sciences are the arbiter in metaphysical matters, with some seeing all of the various sciences as relevant, while others favor the natural sciences over the social sciences, and still others favor an even more restrictive view according to which fundamental physics is the only relevant science for deciding issues of existence. Even if one were to resolve the issue of which sciences are to decide metaphysical matters, however, metaphysics is not the whole of philosophy. Precisely what methodological conclusions one might draw about other areas of philosophy once one figures out which science answers metaphysical questions is anything but clear.

It would be a mistake to conclude, however, that a philosopher's commitment to naturalism has no implications for proper method in philosophy. Various versions of naturalism have profound methodological implications, and many philosophers go out of their way to make clear the connections between their naturalistic commitments and their preferred methodology. Little is accomplished, however, by giving some stipulative account of what naturalism "really" is, and then drawing out the methodological consequences of the favored stipulation. There is a broad family of views which are recognizably naturalistic in that they all see deep connections between science and philosophy, with philosophy owing a good deal of deference to science in matters of substance and/or method. We will do best in trying to understand the various methodologies which naturalism gives rise to by looking at the range of methodological views and practices which may be found among contemporary naturalists.

It is worth pointing out as well that the methodological and metaphysical views of naturalism are not, of course, unrelated (Kornblith, 1994). The idea that science is owed great deference is due to a view about its reliability, and any view about the reliability of science will inevitably bring with it views about both methodology and metaphysical commitments. Some authors put more emphasis on one of these features rather than the other, or, alternatively, see one of these as more fundamental than the other. One might, however, see them as two sides of the same coin.

There is one further issue which needs to be addressed here before moving on, and that is the relationship between naturalism and views about the a priori. Under Quine's influence, naturalism had, at least for a long time, been widely thought to include a rejection of the a priori, either by endorsing the view that all knowledge is empirical, or, more faithfully capturing the point of Quine's critique, rejecting the significance of the very distinction between the a priori and the a posteriori. It is certainly true that a great many naturalists are still sympathetic to such a view (see, e.g. Devitt, 2005). Nevertheless, there are a number of naturalists who have defended accounts of the a priori within the confines of a naturalistic worldview (see, e.g. Kitcher, 1980; Rey, 1993, 1998; Goldman, 1999; Antony, 2004.) Indeed, consider the following possibility, envisioned by Georges Rey, Alvin Goldman, and Louise Antony. Suppose we were to accept a reliabilist theory of justified belief, and, suppose too, that there should be a mental module which is hardwired with some basic logical principles sufficient to allow for the reliable generation of logical theorems without requiring any sensory input at all. On the reliabilist view, the resulting beliefs would be justified, and, since they have no sensory input, they might reasonably be regarded as a priori justified. There seems nothing in a naturalistic epistemology which must lead us to reject such a view, and there is, in addition, some psychological evidence that the claim about such a mental module might well be true (Rips, 1994). Given the current disagreement among naturalists, then, about the tenability of the a priori/a

posteriori distinction, we would do well not to identify naturalism with a commitment to its rejection.

2. Using Science to Address Philosophical Questions

Consider, for example, Jerry Fodor's work on the language of thought (Fodor, 1975; 2008) and on modularity (Fodor, 1983). In each of these works, Fodor examines a wide range of claims made within a successful science, and he argues that these claims have unnoticed implications for topics in the philosophy of mind. Thus, Fodor has argued that cognitive scientists have long been talking about psychological processing in ways which presuppose the existence of a representational medium in which that processing takes place, and this representational medium must have a number of important properties in common with natural languages. Hence, he argues, we should be committed to the existence of a "language of thought." Similarly, Fodor argues that a number of very successful research programs in cognitive science presuppose that the mind has a certain large-scale structure—that it is composed, in part, of subject-matter specific modules which process information in a way which is encapsulated from the larger body of information stored within memory.

There is a common argumentative structure to these works. A body of literature is identified within a well-established science. Insofar as the science has a progressive history—that is, it has an established track-record of ever greater predictive and explanatory success—we have evidence that the theories of these sciences are at least approximately true.[2] And to the extent that we have evidence for the approximate truth of these theories, we thereby have evidence as well for the truth of their previously unnoticed implications (see also Boyd, 1983). So to the extent that Fodor has rightly identified presuppositions of going theories in the cognitive sciences, and to the extent that the cognitive sciences exhibit a progressive history, we thereby have reason to believe that these presuppositions—that there is a language of thought, and that the mind has a modular structure—are true. Fodor then goes on to argue that these claims have important implications for a wide range of traditional philosophical problems: about the relation between thought and language; about rationalism and empiricism; and so on.

On this way of approaching philosophical questions, the results of the sciences have rather direct implications for issues with which philosophers have long been concerned. Philosophy, on such an approach, does not have some distinctive methodology which sets it apart from the sciences. Instead, advances in the sciences allow us to address traditional philosophical problems. When Plato and Descartes, for example, attempted to understand

[2] Some will prefer to make this point, and the points which follow in this paragraph, without adverting to truth. Thus, one might claim that we are warranted in accepting our most successful scientific theories, and such warrant extends to any philosophical claims which flow from them. I prefer the formulation in the text, not only for ease of exposition, but because I believe that the commitment to scientific realism it embodies provides us with the most defensible form of naturalism.

the structure of the mind, they could not, of course, turn to the cognitive sciences to address those issues. The methods which they employed, however, should not be seen as definitive of legitimate philosophical inquiry. Instead, we should simply recognize that Plato and Descartes were unfortunate enough to have been born too early to have available the sorts of tools and theoretical frameworks which were essential for approaching this particular subject-matter in a successful manner. It is not that some single datum serves to address long-standing philosophical problems. It was not, for example, the result of a particular reaction-time experiment which led to the insights Fodor offers. If this approach is correct, however, philosophical insight is nevertheless a direct product of the development of successful scientific theories. There is no autonomous philosophical methodology, on this way of approaching matters, for the solution to philosophical problems comes from science itself.

Fodor offers us a particularly pure example of this naturalistic methodology, but he is by no means idiosyncratic. While philosophers of mind in the first half of the twentieth century rarely turned to experimental psychology as a way of addressing philosophical problems, philosophy of mind since the nineteen sixties has shown an ever greater degree of empirical engagement. Philosophical questions about consciousness, perception, mental representation, decision, and action regularly engage with the empirical literature on these topics. One may, to be sure, see the empirical literature as presenting certain constraints on philosophical theories without going so far as to see the solution to philosophical questions as a direct product of scientific work. There can be little doubt, however, that the entire texture of work in the philosophy of mind has changed dramatically since the first half of the twentieth century, with empirical work playing a larger and larger role in philosophical theorizing.

Philosophy of mind has played a leading role here, but it is by no means the only area in which scientific theories are mined for insight into philosophical issues. It is worth asking, however, why it is that philosophy of mind has played the role it has. If we compare, for example, philosophy of mind with work in ethics, there is an extremely striking contrast. Most of the work which goes on in ethical theory makes little if any use of scientific literature, but, of course, this is not terribly surprising. While the mind is equally a subject of study by philosophers and experimental psychologists, there does not seem to be any area of scientific study which is occupied with work on, for example, the good and the right. An ethicist who was impressed by the success of Fodor's work in the philosophy of mind might understandably think that there is simply no way in which one might take a similar approach to ethics. It is thus worth asking—even if we think that Fodor's approach to questions in philosophy of mind is especially fruitful—the extent to which this approach can be brought to bear on other areas within philosophy.

Work in the philosophy of science has, unsurprisingly, drawn on scientific results in much the way that Fodor and others have in the philosophy of mind. Just as philosophy of mind and the cognitive sciences have a common subject-matter, philosophy of physics, biology, and so on share a subject-matter with the sciences they examine. Philosophical work on space and time has long been shaped by work in physics (see, for example, Sklar, 1974.) More than this, physics has often been viewed as a source of insights into general questions in metaphysics (e.g. Ladyman, Ross, Spurrett, and Collier, 2007; Maudlin, 2007.) Similarly, work in philosophy of biology draws on the substance of biological theories (e.g.

Kitcher, 2003; Sober, 1993, 2000.) The days in which philosophy of science was dominated by abstract discussions of the logic of science in the absence of detailed discussion of particular scientific theories is largely over.[3]

Similarly, there is work in epistemology which draws on results in the cognitive sciences as a way of understanding the nature of knowledge. Alvin Goldman has been instrumental here in getting epistemologists to pay attention to the import of empirical work for understanding epistemological issues (see, for example, Goldman 1986, 1992). Ruth Millikan (1993), Peter Godfrey-Smith (1996), and Kim Sterelney (2003) have set epistemology in a biological framework, and Hilary Kornblith (2002) has attempted to draw epistemological conclusions from work in cognitive ethology. There has also been a great deal of work, initially inspired by Herbert Simon's work on bounded rationality (Simon, 1969, 1982), and Amos Tversky and Daniel Kahneman's investigations of heuristics and biases in human inference (see Nisbett and Ross, 1980; Kahneman, Slovic, and Tversky, 1982; and Gilovich, Griffin, and Kahneman, 2002), which seeks to develop an epistemology which takes account, not only of human limitations, but of the ways in which our cognitive successes are actually achieved in practice (see, e.g. Kornblith, 1993; Stein, 1996; Gigerenzer et al., 1999; Morton, 2013.)

Although there is arguably no science of the good and right, work in ethics has, in fact, been influenced in important ways by the sciences, and many have drawn moral conclusions from empirical work. There is a good deal of work, not only on the ways in which evolution might explain the rise of co-operative behavior, but on ways in which evolution might shed light on morality itself (see, e.g. Gibbard, 1990; Skyrms, 2003; Sober and Wilson, 1998). Some moral realists (see, e.g. Boyd, 1988) have attempted to show just how it is that we might understand moral properties as objects of empirical investigation.

Naturalists who favor the kind of approach to philosophical questions described here are often faced with the question of whether what they are doing is really philosophy. Fodor's response to this kind of concern is, I believe, illustrative of the naturalistic approach to these issues:

> Some of the arguments I have on offer are patently philosophical; some turn on experimental and linguistic data; many are methodological; and some are just appeals to common sense. That there is no way of talking that is comfortable for all these sorts of dialectic is part of what makes doing cognitive science so hard. In the long run, I gave up; I've simply written as the topics at hand seemed to warrant. If it doesn't sound exactly like philosophy, I don't mind; as long as it doesn't sound exactly like psychology, linguistics, or AI either.
>
> (Fodor, 1998, viii)

3. Experimental Philosophy

Quite a different methodological approach to philosophical questions, but one no less naturalistic, is illustrated by work in experimental philosophy (see Knobe and Nichols, 2008;

[3] For an exceptionally illuminating discussion of the ways in which naturalistic themes have played a role in the development of philosophy of science in the late twentieth century, see Kitcher, 1992.

Horvath and Grundmann, 2012; Alexander, 2012, esp. chapter 4). Perhaps the best way to understand work in this area is by seeing what it shares with, and how it departs from, traditional conceptual analysis.

Philosophers engaged in conceptual analysis, as it has traditionally been practiced, have attempted to understand important philosophical concepts by way of examining our intuitions about hypothetical cases. Thus, for example, if we wish to understand the concept of knowledge, we might examine our intuitions about cases in which a subject arrives at a certain belief in carefully specified circumstances to see whether we have the intuition that a subject so situated would count as having knowledge. An analysis of the concept of knowledge would attempt to systematize these intuitions, providing necessary and sufficient conditions for knowledge in such a way as to square with the intuitions about individual cases.[4] Similarly, one might seek to provide philosophical analyses of the concept of the good, or the right, or of various mental states, or freedom of the will, and so on.

Many philosophers who engage in conceptual analysis of roughly this sort see themselves as providing analyses, not of their own, private, potentially idiosyncratic concepts, but rather of concepts which are widely shared. They speak of analyzing "our" concepts, or of "folk concepts" (see, e.g. Jackson, 1998). It is striking, however, that philosophers in this tradition attempt to provide such analyses on the basis, typically, of their own intuitions about hypothetical cases. And this seems to raise a problem: Why should one think that one's own intuitions about cases are a good indication of the intuitions of the folk? If one were genuinely interested in the intuitions which people within a given speech community share, wouldn't it be important to engage in some sort of empirical investigation of the intuitions within that community, rather than simply rely on one's own intuitions, in the absence of any real information about the intuitions of others? (Jackson briefly addresses this issue: 1998, 36–7.)

It is this sort of concern which prompts the experimental turn in philosophy. By using carefully designed questionnaires, experimental philosophers have examined the intuitions of large groups of individuals, with an eye toward discovering, not only the extent to which various intuitions are shared, but also the nature of the processes by which those intuitions are produced. Thus, for example, Jonathan Weinberg, Shaun Nichols, and Stephen Stich (2001) found that the Gettier intuition—that one might have justified, true belief which is not knowledge—was widespread among those of European descent, but far less common among East Asians. Stacy Swain, Joshua Alexander, and Jonathan Weinberg (2008) found that a certain series of examples widely appealed to in order to show a defect in reliabilist accounts of justification would elicit exactly the opposite intuition—one which would support reliabilist accounts—when the examples were presented in reverse order. (This result has been disputed by Nagel (forthcoming) and Wright (2010).) This kind of work, sometimes referred to as "the negative program in experimental philosophy," is used to cast doubt on the status of intuitions used as data in traditional conceptual analyses. If

[4] A proper analysis need not conform to every such intuition. Some intuitions may be explained away, rather than explained by, a proposed analysis. Nevertheless, intuitions about hypothetical cases are seen as checkpoints for philosophical theorizing, much as observational data are viewed as checkpoints for scientific theorizing.

our intuitions about when we have knowledge or justification are influenced by our ethnicity, or gender, or social class, then they surely provide little insight into how we ought to go about arriving at our beliefs. If the examples used to undermine reliabilism would produce a different intuition if only they were presented in a different order, then there is little reason to see them as presenting a serious challenge to reliabilist accounts. Experimental work of this sort may be used to provide a check on the kind of data which is widely regarded—especially among traditional conceptual analysts—as the proper evidential base for philosophical theorizing.

But this is not the only sort of work done by experimental philosophers. By understanding the source of various intuitions, experimental philosophers have sought to understand how it is that people ordinarily think about a variety of philosophically interesting topics. For example, Joshua Knobe (2003) and Philip Pettit and Joshua Knobe (2009) found that people's intuitions about whether an action is intentional are strongly influenced by whether the action is seen as moral or immoral. While one might have thought that questions about whether an action is intentional are entirely independent of the moral status of the action, this work strongly suggests that this is not at all how people typically think about such matters. Studies of this sort—and experimental philosophers have examined intuitions about a wide variety of philosophically interesting topics—may provide substantial insight into our folk concepts. Because the results achieved by way of this experimental work are often quite different from what one might otherwise have expected, and because these results are often at variance with the kinds of conceptual analyses produced by more traditional armchair means, this work has commanded wide attention.

What is the philosophical significance of this kind of work? Not surprisingly, experimental philosophers have often faced the very kind of criticism addressed to the empirically minded philosophers discussed in section 2.1. This is all very interesting, some will say, but what does it have to do with philosophy? (See, for example, the critical essays in Horvath and Grundmann, 2012, also Cappelen, 2012.) For experimental philosophers, however, the answer to this is quite clear: this kind of work allows us to carry out philosophical investigations by non-traditional means. At a minimum, as suggested above, it allows us an independent check on the reliability of the intuitions upon which more traditional philosophers rely. More than this, it may be seen as carrying on conceptual analysis by more careful and empirically responsible techniques. If we want to know what our folk concepts are—as many armchair conceptual analysts say they do—then experimental philosophers will argue that traditional methods are no more well suited to this task than armchair thinking about physics is to understanding the fundamental laws of nature. The way in which we think about philosophical topics, and the character of our folk philosophical concepts, is simply not available to introspection, nor can it be revealed by way of armchair theorizing. Experimental philosophy is often viewed by its critics as an attempt to stop doing philosophy and, instead, talk about something else. But as experimental philosophers see it, this thoroughly misunderstands the enterprise. Experimental philosophy doesn't reject traditional philosophical questions about our folk concepts and replace them with questions about something else. What it does, instead, is propose new and better methods for addressing the very questions about our folk concepts which philosophers have long been asking.

4. Naturalism and Armchair Philosophy

The naturalistic methodologies we have examined thus far are frankly dependent on evidence unavailable from the armchair. Philosophers who favor these methodologies have offered extended critiques of armchair philosophizing. (For a particularly pointed critique, see Cummins, 1998) Nevertheless, there are naturalists who defend the legitimacy of armchair methods, and there are a several distinct approaches which have been offered in furtherance of such a defense.

Alvin Goldman (2007) has argued in favor of a view of philosophy as conceptual analysis, where concepts are seen as psychologically real entities which play a causal role in producing various sorts of behavior. For example, if we want to explain a subject's ability to recognize the difference between animals which are mammals and those which are not, we need to appeal, on this view, to a feature of the subject's psychology: the subject's concept of a mammal. Features of the subject's concept explain how it is that the subject is able to recognize mammals and distinguish them from non-mammals, and it explains the inferences the subject makes on discovering, for example, that a particular animal is a mammal. Similarly, subjects may have concepts of knowledge, justification, belief, desire, the good, the right, justice, and so on, and these philosophically interesting concepts play an important role in the classifications that individuals make as well as the inferences which they draw. When subjects are presented with hypothetical cases, their concepts play a causal role in producing the judgments they make about those cases. Thus, when a subject is presented with a Gettier case, the subject's judgment that the case involves justified, true belief, but not knowledge—assuming the subject makes such a judgment—is produced by way of the subject's concept of knowledge. On this view, then, we may probe a subject's intuitions by way of hypothetical cases, and we may use these intuitions as evidence for theorizing about the concept which played a role in producing the intuition. Philosophy, on this view, is in the business of explicating our concepts, and Goldman offers a psychological theory about why it is that appealing to our intuitions—and thus armchair methods in philosophy—may offer a path to successful philosophical theorizing.

Of course, the route from concept to intuition may not proceed smoothly, for there may be psychological factors which interfere with the natural expression of the concept, and Goldman is quite frank about this. At the same time, Goldman suggests that there is good reason to be optimistic that armchair methods will be largely successful in delivering central features of our concepts, even if we may need to engage in more elaborate experimental methods of the sort favored by experimental philosophers in order to refine our conceptual analyses. Thus, although Goldman offers a defense of armchair methods in philosophy, he allows that experimental approaches will play some role in philosophy as well.

A number of philosophers have attempted to offer experimental evidence in favor of the legitimacy of armchair methods, thereby bolstering the case which Goldman makes in favor of a traditional philosophical methodology. Jennifer Nagel (2012) has argued that intuitions about hypothetical cases may well be the product of a modular faculty which issues in judgments which are very likely to be true, and thus these judgments may serve as appropriate data in constructing philosophical theories. A similar suggestion has been made by Joseph Shieber (2012). Although Nagel and Shieber do not commit themselves to

Goldman's view that such theorizing is targeted at features of our concepts, they share with Goldman the view that traditional armchair methods in philosophy are largely legitimate. Since a good deal of philosophical theorizing seems to have employed armchair methods, anyone who thinks that philosophical theories have often been a source of illumination is likely to see this result as making sense of the enterprise of philosophy in a way that those who challenge traditional methods may not.

It is a striking feature of the accounts offered by Goldman, Nagel, and Shieber that they all offer substantive psychological theories, and a variety of empirical evidence in their favor, as a way of defending traditional armchair methods. The naturalistic character of their approach is thus clear, in spite of the fact that, in the end, the methodologies they favor look very similar to the approaches of philosophers who are utterly opposed to naturalism, both in substance and in method.

A very different sort of defense of armchair methods, yet still from within a naturalistic perspective, is found in the Canberra Plan (see Braddon-Mitchell and Nola, 2009) The central idea behind this approach was originally formulated by David Lewis (1970, 1972). Lewis offers an account of how to define theoretical terms, with special reference to psychological vocabulary. The suggestion, roughly, is as follows. If we wish to give a definition for some psychological term—say, "pain"—we may begin by listing all of the platitudes about it. Thus, we will have that it tends to be caused by injury; that it tends to cause certain sorts of behavior; and so on. We may now offer a definition of "pain" as that state which has all of these properties, or, alternatively, which has most of these properties or comes close to having most of these properties. (I will largely ignore these qualification in what follows for ease of exposition.) No doubt many of the platitudes about pain relate it to other mental states, but we can define all of the mental state terms at once by generalizing this approach in a manner suggested by Ramsey (1990). If we wish to know what pain is, we may now engage in an empirical investigation to determine what state in the world actually satisfies these conditions.

We may use this approach to define philosophical terms generally, and thus the task of providing philosophical analyses comes down to the armchair project of laying out the various platitudes about, say, pain, or freedom, or persons, and constructing the appropriate existentially quantified sentence which identifies the referent of the term with whatever it is that (approximately) satisfies (most of) the platitudes.

As Lewis presented this, one of the motivations for the project was that it allowed us to see how the referents of mental state terms might be nothing more than certain physical states. This way of defining theoretical terms thus allowed for the domestication of the mental: the place of the mind in the physical world is thereby made plain. More than this, when the project is generalized to allow for the definition of philosophical terms across the board in terms of their relations to thoroughly unproblematic entities, states, and processes, this way of proceeding allows for a more thoroughgoing project of domestication: just as this project makes plain the place of the mind in the physical world, it may be used to make plain the place of other prima facie problematic entities, states, and processes in a thoroughly unproblematic world. The appeal of this project to those who are attracted to a naturalistic metaphysics was immediately clear.

It is for this reason that many have followed Lewis in adopting a certain style of armchair methods in the service of a naturalistic picture of the world (see Jackson, 1998 and

the essays in Braddon-Mitchell and Nola, 2009.) While there is nothing intrinsically naturalistic about the project of constructing definitions out of platitudes, this method can be recruited to serve the purposes of defending a naturalistic metaphysics, and this is precisely how it plays out in the hands of Lewis and the Canberra Planners.

There are thus a number of different routes by which naturalists may be led to favor armchair methods in philosophy.

5. Conclusion

There are thus a number of competing methodologies which naturalists have favored, some of which look very little different from the methods of philosophers who have no sympathy with naturalism, while others mark a distinctive break with traditional philosophical methods.[5]

References

Alexander, Joshua, *Experimental Philosophy: An Introduction*, Polity Press, Boston, 2012.

Antony, Louise, "A Naturalized Approach to the A Priori," *Philosophical Issues*, 14(2004), 1–17.

Boyd, Richard, "On the Current Status of Scientific Realism," *Erkenntnis*, 19(1983), 45–90.

Boyd, Richard, "How to Be a Moral Realist," in Geoffrey Sayre-McCord, ed., *Essays on Moral Realism*, Cornell University Press, Ithaca, New York, 1988, 181–228.

Braddon-Mitchell, David and Nola, Robert, *Conceptual Analysis and Philosophical Naturalism*, MIT Press, Cambridge, MA, 2009.

Cappelen, Herman, *Philosophy without Intuitions*, Oxford University Press, Oxford, 2012.

Cummins, Robert, "Reflections on Reflective Equilibrium," in Michael DePaul and William Ramsey, *Rethinking Intuition: The Psychology of Intuition and its Role in Philosophical Inquiry*, Rowman and Littlefield, Lanham, MD, 1998, 113–27.

Devitt, Michael, "There is no A Priori," in Matthias Steup and Ernest Sosa, eds., *Contemporary Debates in Epistemology*, Blackwell, Oxford, 2005, 105–15.

Fodor, Jerry, *The Language of Thought*, Crowell, New York, 1975.

Fodor, Jerry, *The Modularity of Mind*, MIT Press, Cambridge, MA, 1983.

Fodor, Jerry, *Concepts: Where Cognitive Science Went Wrong*, Oxford University Press, Oxford, 1998.

Fodor, Jerry, *LOT 2: The Language of Thought Revisited*, Oxford University Press, Oxford, 2008.

Gibbard, Allan, Wise Choices, Apt Feelings: A Theory of Normative Judgment, Harvard University Press, Cambridge, MA, 1990.

Gigerenzer, Gerd, Todd, Peter, and the ABC Research Group; *Simple Heuristics that Make Us Smart*, Oxford University Press, Oxford, 1999.

[5] I am grateful to the editors, an anonymous referee, and Brian Leiter for helpful comments on a previous draft of this chapter.

Gilovich, Thomas, Griffin, Dale, and Kahneman, Daniel, eds., *Heuristics and Biases: The Psychology of Intuitive Judgment*, Cambridge University Press, Cambridge, 2002.
Godfrey-Smith, Peter, *Complexity and the Function of Mind in Nature*, Cambridge University Press, Cambridge, 1996.
Goldman, Alvin, *Epistemology and Cognition*, Harvard University Press, Cambridge, MA, 1986.
Goldman, Alvin, *Liaisons: Philosophy Meets the Cognitive and Social Sciences*, MIT Press, Cambridge, MA, 1992.
Goldman, Alvin, "A Priori Warrant and Naturalistic Epistemology," *Philosophical Perspectives*, 13(1999), 1–28.
Goldman, Alvin, "Philosophical Intuitions: Their Target, Their Source, and Their Epistemic Status," *Grazer Philosophische Studien*, 74(2007), 1–26.
Horvath, Joachim and Grundmann, Thomas, eds., *Experimental Philosophy and its Critics*, Routledge, London, 2012.
Jackson, Frank, *From Metaphysics to Ethics: A Defence of Conceptual Analysis*, Oxford University Press, Oxford, 1998.
Kahneman, Daniel, Slovic, Paul and Tversky, Amos, eds., *Judgment under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge, 1982.
Kitcher, Philip, "A Priori Knowledge," *Philosophical Review*, 89(1980), 3–23.
Kitcher, Philip, "The Naturalists Return," *Philosophical Review*, 101(1992), 53–114.
Kitcher, Philip, *In Mendel's Mirror: Philosophical Reflections on Biology*, Oxford, University Press, Oxford, 2003.
Knobe, Joshua, "Intentional Action and Side-Effects in Ordinary Language," *Analysis*, 63(2003), 190–3.
Knobe, Joshua and Nichols, Shaun, eds., *Experimental Philosophy*, Oxford University Press, Oxford, 2008.
Kornblith, Hilary, *Inductive Inference and its Natural Ground: An Essay in Naturalistic Epistemology*, MIT Press, Cambridge, MA, 1993.
Kornblith, Hilary, "Naturalism: Both Metaphysical and Epistemological," *Midwest Studies in Philosophy*, XIX(1994), 39–52.
Kornblith, Hilary, *Knowledge and its Place in Nature,* Oxford University Press, Oxford, 2002.
Ladyman, James, Ross, Don, with Spurrett, David, and Collier, John, *Every Thing Must Go: Metaphysics Naturalized*, Oxford University Press, Oxford, 2007.
Lewis, David, "How to Define Theoretical Terms," *Journal of Philosophy*, 67(1970), 427–66.
Lewis, David, "Psychophysical and Theoretical Identifications," *Australasian Journal of Philosophy*, 50(1972), 249–58.
Maudlin, Tim, *The Metaphysics within Physics*, Oxford University Press, Oxford, 2007.
Millikan, Ruth, *White Queen Psychology and Other Essays for Alic*e, MIT Press, Cambridge, MA, 1993.
Morton, Adam, *Bounded Thinking: Intellectual Virtues for Limited Agents*, Oxford University Press, Oxford, 2013.
Nagel, Jennifer, "Intuitions and Experiments: A Defense of the Case Method in Epistemology," *Philosophy and Phenomenological Research*, 85(2012), 495–527.
Nisbett, Richard and Ross, Lee, *Human Inference: Strategies and Shortcomings of Social Judgment*, Precntice-Hall, Englewood Cliffs, NJ, 1980.
Pettit, Dean and Knobe, Joshua, "The Pervasive Impact of Moral Judgment," *Mind and Language*, 24(2009), 586–604.

Quine, W. V. O., *Ontological Relativity and Other Essays*, Columbia University Press, New York, 1969.
Ramsey, Frank, *Philosophical Papers*, D. H. Mellor, ed., Cambridge University Press, Cambridge, 1990.
Rey, Georges, "The Unavailability of What We Mean I: Quine, Fodor and LePore," *Grazer Philosophische Studien*, 46(1993), 61–101.
Rey, Georges, "A Naturalistic A Priori," *Philosophical Studies*, 92(1998), 25–43.
Rips, Lance J., *The Psychology of Proof: Deductive Reasoning in Human Thinking*, MIT Press, Cambridge, MA, 1994.
Sellars, Wilfrid, *Science, Perception and Reality*, Routledge and Kegan Paul, New York, 1963.
Shieber, Joseph, "A Partial Defense of Intuition on Naturalist Grounds," *Synthese*, 187(2012), 321–41.
Simon, Herbert, *The Sciences of the Artificial*, MIT Press, Cambridge, MA, 1969.
Simon, Herbert, *Models of Bounded Rationality*, vols. 1 and 2, MIT Press, Cambridge, MA, 1982.
Sklar, Lawrence, *Space, Time and Spacetime*, University of California Press, Berkeley, 2007.
Skyrms, Brian, *The Stag Hunt and the Evolution of Social Structure*, Cambridge University Press, Cambridge, 2003.
Sober, Elliott, *The Nature of Selection: Evolutionary Theory in Philosophical Focus*, University of Chicago Press, Chicago, 1993.
Sober, Elliott, *Philosophy of Biology*, 2nd edition, Westview Press, Boulder, CO, 2000.
Sober, Elliott and Wilson, David Sloan, *Unto Others: The Evolution and Psychology of Unselfish Behavior*, Harvard University Press, Cambridge, MA, 1998.
Stein, Edward, *Without Good Reason: The Rationality Debate in Philosophy and Cognitive Science*, Oxford University Press, Oxford, 1996.
Sterelney, Kim, *Thought in a Hostile World: The Evolution of Human Cognition*, Blackwell, Oxford, 2003.
Swain, Stacey, Alexander, Joshua, and Weinberg, Jonathan, "The Instability of Philosophical Intuitions: Running Hot and Cold on Truetemp," *Philosophy and Phenomenological Research*, 76(2008), 138–55.
Weinberg, Jonathan, Nichols, Shaun, and Stich, Stephen, "Normativity and Epistemic Intuitions," *Philosophical Topics*, 29(2001), 429–460.
Wright, Jennifer Cole, "On Intuitional Stability: The Clear, the Strong, and the Paradigmatic," *Cognition*, 115(2010), 491–503.

CHAPTER 9

METHOD IN ANALYTIC METAPHYSICS

DANIEL NOLAN

1. Introduction

There is no one agreed method in contemporary metaphysics. Methodological disputes in contemporary metaphysics run deep: each of the main methods discussed in this article will be denounced as worthless or pernicious by at least some writers. Despite this, generalizations about contemporary methods are possible, provided that these generalizations are not treated as describing every practising metaphysician's work. The vast majority of contemporary metaphysicians in the broadly "analytic" tradition will identify some of the methods discussed in this article as important in their own work. Or at least that is one aim of this piece.

The first half of this chapter will discuss five important sources of constraints on metaphysical theorizing: linguistic and conceptual analysis; consulting intuitions; employing the findings of science; respecting folk opinion; and applying theoretical virtues in metaphysical theory choice such as preferring simpler theories, or preferring more explanatory theories. The second half of this chapter will discuss four other topics of particular methodological interest in metaphysics: the role of formal methods in metaphysics; the role of metaphysical communities, traditions, and the place of the history of metaphysics in contemporary work; the issue of whether metaphysics should be seen as an enterprise that will yield knowledge of metaphysical matters or whether it should have more modest epistemic goals; and the question of how much of metaphysics is a priori.

2. Linguistic and Conceptual Analysis

Analytic philosophy, narrowly defined, saw the main task of philosophy as the analysis of language, and held that many traditional philosophical problems could be solved, or

dissolved, by careful attention to the language in which they were framed. From this narrow point of view, philosophers doing metaphysics could be doing one of three things. They could be showing how analysis of the language used in traditional metaphysical questions enabled us to see what the solutions were, or that the questions were ill-posed or meaningless. (Good.) They could be engaged in trying to discover the answers to questions which were the result of linguistic mistakes by treating them as substantive questions. (Bad.) Or they could be trying to discover the answer to general questions about the world, of the sort that are genuinely substantive, using methods only suitable for resolving linguistic confusions. (Also bad. Questions like that, if they were to be answered at all, were to be answered by physics or some other inquiry.)

This outline is at best an over-generalization, and at worst a caricature, though there seem to have been some philosophers whose attitude to metaphysics could be summed up in this way. That sort of analytic philosophy, narrowly conceived, was rightly seen as hostile to metaphysics (and "metaphysics" sometimes was used as a term of abuse for the second of the tasks above: see Ayer 1936 ch. 1, for example). Contemporary analytic philosophy is rarely "analytic philosophy" in the narrow sense above: many philosophers today think that good philosophy can be much more than linguistic analysis, and that it draws on sources of evidence other than observations of the functioning of language and thought. But the use of analysis of language is still one of the tools in the analytic philosophy toolkit. Related to the analysis of language, many philosophers still value "conceptual analysis". Some use this as a synonym for the analysis of language (see Jackson 1998 pp. 33–4), but for many the difference between conceptual analysis and linguistic analysis is the difference between analysing language and analysing mental content, that is, the content of *concepts*.

So, what could an analysis of either language or our concepts have to tell us about metaphysical topics such as causation, the nature of space and time, free will, the existence of abstract objects, and so on? One might hope to establish many interesting things: that effects must never temporally precede their causes (Black 1956 esp. p. 55), or that there being as many Fs as Gs is sufficient to ensure that there is an abstract entity, a number, that "numbers" the Fs. (See Wright 1999 for discussion.) Sometimes linguistic analysis is treated, not as something to be done entirely by reflection on one's understanding of the words, but along the lines of an explanatory linguistic hypothesis. A familiar example is Davidson's conjecture (Davidson 1967) that the best theory of adverbs involves quantification over events: this is meant to explain how "Jim slowly ate a banana in the kitchen" entails "Jim ate a banana in the kitchen" and "Jim ate a banana", for example. Davidson's analysis, if accepted, would seem to support an ontology of events, and so is of metaphysical interest as well as being of interest to the philosopher of language.

In the contemporary climate, a metaphysician attempting to establish significant metaphysical conclusions through analysis either of concepts or language is likely to face several challenges. In the case of concepts, one challenge is to come up with a plausible story about concepts such that examination of them (or "analysis" of them, whatever that is) can yield useful information. One part of this is the question of whether concepts, as opposed to, for example, beliefs, encode propositional information at all. Another part of the challenge is an epistemic one: why ought we rely on information somehow encoded in our concepts? A third part of the story will be needed for those who think that concepts can tell us *how*

the world is and not merely (e.g.) how we represent it: how can concepts reflect the world, and can they do this in a way that enables us to discover things about the world through examination of our concepts? (Jenkins 2008 Part 1 is an interesting recent attempt to offer an epistemology of knowledge through conceptual reflection that tackles the second and third problems mentioned.)

Similar challenges arise if one wants to come to interesting metaphysical conclusions through analysis of language. When linguistic analysis purports to discover truths through examination of language, this is not in the ordinary way of seeing what people say and whether, for example, they knew what they asserted. The defender of linguistic analysis needs a story about how else words or sentences contain information that can be extracted through "analysis". They need a story about why the information so extracted can be relied upon (e.g. even if some word use presupposes something metaphysical, why does uncovering that presupposition yield justification for believing the metaphysical presupposition?) Finally, if examination of words and sentences is to yield information about how the world is, metaphysically speaking, rather than, for example, how competent word users tend to take it to be, how can this happen, and how can it happen in such a way that we can be justified in taking the non-linguistic world to be a certain way through examination of language?

To say these challenges arise for linguistic analysis is not to say they cannot be answered. (I offer a tentative response to some of these challenges in Nolan 2009, for example, at least if what I am defending counts as "linguistic analysis" at all. For a different and more ambitious defence of conceptual/linguistic analysis, see Jackson 1998). One problem with hoping to extract anything very interesting from conceptual or linguistic analysis is that anything so extracted is likely to be controversial: and what is more, controversial among those who apparently competently deploy the concept or vocabulary in question. Presumably even experts can make conceptual or linguistic mistakes, but it is dialectically awkward, to say the least, to accuse those who hear the "analysis" and continue to disagree as either being incompetent with the concept or term involved, or suffering a persistent performance error. This problem can be avoided in one of two ways. One is to insist that recognition of the correctness of an "analysis" requires considerably more than competent possession and successful use of the concept or word involved. (This is plausibly what happens when an "analysis" flows from a mature linguistic theory or a cognitive science theory of the behaviour of a particular concept.) The other is to aim for "analysis" that is as uncontroversial as is feasible. Analyses which deliver weak generalizations, or disjunctions of interesting disjuncts, might well be acceptable to a wider range of competent users than the exciting proposed analyses of the heyday of analytic philosophy.

"Causes must always precede their effects" is a controversial claim, and it is especially controversial that it is a result of conceptual or linguistic analysis. "Causes mostly, actually, around here, precede their effects" is probably not entirely uncontroversial, but I suspect it is both much less controversial per se, and less controversial that it might somehow be produced by conceptual or linguistic analysis. An analysis which yields five necessary conditions to be a mind, for example, will be more controversial than one that says it is necessary to be a mind to have at least three of those five conditions. (The second analysis will be entailed by the first, so it had better not be

any *more* controversial!) Weakening analyses may remove some of the point of seeking them in the first place, but weak constraints are still significantly better than no constraints at all.

Another challenge that many linguistic and conceptual analyses face is that when these analyses deliver demanding conditions for something to answer to a metaphysical concept, they often bring with them the threat that it will turn out that nothing in our world falls under that concept. To take one example: suppose I initially thought that there could be cases of effects temporally preceding their causes (because of apparently consistent time travel stories, tachyon theories, models of spacetime with loops, or whatever). Let us suppose that conceptual or linguistic analysis then showed conclusively that any relation that fell under our concept CAUSE (or word "cause") must always go from past to future. It is open to me to then conclude that the relation that standardly holds in this world between, for example, swinging feet and flying footballs, sunlight and plant growth, etc. etc. is not causation at all, but only causation*: a relation just like causation is supposed to be, except that it sometimes goes from the future to the past.

How serious this challenge is may vary from case to case. Sometimes we are sufficiently certain that there are examples of whatever it is we are analysing (causation, location in a region of space, properties, or whatever) that we might be sure that e.g. there *really is* causation, and not just causation*. Though a critic might wonder whether we should still be sure that there is causation if it turns out the concept of causation is more demanding than we initially thought.

Even if conceptual and linguistic analysis is ill-suited to establishing important metaphysical claims on its own, it may still have an important place in metaphysical inquiry. Careful thinking about what we say and think might reveal tensions in our initial opinions about a topic, even if it is hard to say whether those tensions somehow "flow from the concept" or were confusions or contradictions of some more mundane sort. Careful attention to ordinary ways of talking, or even technical ways of talking, might reveal ambiguity or context-dependence that would trip us up later in inquiry if we were not aware of it. And in determining what ordinary or scientific opinion is (see sections 4 and 5), it is often important to understand the language in which those opinions are expressed. So attention to our thought and talk about metaphysical issues can play an important ground-clearing role, even if it is not the whole of metaphysical inquiry.

One note of caution when reading metaphysicians who offer an "analysis" is in order. There are many things that can go under the name "analysis", and in recent years in particular there has been a revival in metaphysicians offering "real definitions" or "metaphysical analyses" that are not necessarily meant to flow from the meanings of our words or the contents of our concepts. Such analyses might instead convey information about essences, or which entities are constituents of which, or other kinds of metaphysical information. Someone offering this kind of "analysis" might support it in any number of ways, so I am inclined to think that the project of offering this kind of analysis need not be *methodologically* distinctive. But at the very least, these projects may not face the costs and benefits that traditional conceptual/linguistic analysis projects bear.

3. Consulting Intuition

Many metaphysicians see employing intuitions as an important part of their practice. An intuition that particular objects like cups could not be wholly in two places at once; or that Queen Elizabeth could not have been born to different parents (Kripke 1980 pp 110–114); or that it is possible for there to be an otherwise empty space with two qualitatively identical steel spheres in it (Black 1952) are all intuitions that have been relied upon by one author or another.

The name "intuition" might suggest that intuitions are supposed to have a distinctive rational source, and indeed there are defenders of this sort of view: George Bealer (Bealer 1987, 1996) is a contemporary defender of a significant role for epistemically basic rational intuitions in metaphysics and elsewhere. But others see intuitions in a much more deflationary way. Lewis (1983 p. x) considers intuitions to just be opinions, while van Inwagen (1997 p. 309) takes intuitions to merely be beliefs, or at least tendencies to accept certain beliefs. If intuitions have nothing distinctive about them other than that they are opinions, for example, then they are unlikely to have a distinctive methodological role. (That is not to say that intuitions will be worthless, either, any more than other beliefs we have are worthless.)

For our purposes, whether intuitions are themselves beliefs, or "seemings" that incline us towards beliefs, is less important than their likely origin and epistemic status. If intuitions are epistemically basic "rational seemings", or alternatively epistemically basic beliefs, and if reliance on them is an important part of metaphysical method, then some effort is justified in improving our intuitional faculty. Presumably some people intuit better than others, or alternatively do better in sorting between true intuitions and other states which can be mistaken for intuitions. Then there is the task of accurately describing the content of one's intuitions. Many philosophers think there is room to honestly misreport what one is intuiting. If that is a risk that is hard to correct for, this is another thing that would require training and practice.

You might expect that there would be self-help works available for improving one's faculty of philosophical intuition, and exercises to improve intuition reporting. There is little in the philosophical literature aimed at this, however. I suspect this is because proponents of intuitional method think the training to become a better intuiter is the same as training to be better at metaphysics in general: reading the writings of others, attempting to construct counterarguments and think of counterexamples, writing metaphysics oneself, and so on. But perhaps those who think there are distinctive exercises of a distinctive faculty at the root of intuition will eventually come up with techniques for improvement more narrowly focused on improving the alleged special faculty.

On the other hand, if intuitions are a grab-bag of seemings, or inclinations to believe, or alternatively a grab-bag of opinions, an important task would seem to be to divide intuitions further in accord with their probable source. Some might well come from tacit conceptual or linguistic competence, and so potentially have the same status as conceptual and linguistic analyses (see section 2). Some might well come from background beliefs, perhaps tacit, about how our world works: these may have sources as humble as everyday experience or well-confirmed scientific discoveries (see section 4). Some might come from

what is taken for granted in one's community of inquiry—perhaps we implicitly learn that some things are so obvious they can be taken for granted even if we do not know any argument for them. (Some prejudices might be like this, and there is no reason to suppose that prejudice is the only place beliefs like these can be communicated.) Perhaps some beliefs or dispositions to believe are somewhat random—we have no a priori guarantee that every belief we have will be properly connected to some identifiable epistemic source. It could of course be, that besides all of these sources of intuitions, there is a leftover core of intuitions that issue from a rational faculty or some equally distinctive source, though it is hard to see how to establish this psychological conjecture when there are so many other plausible sources of how matters seem to us.

As well as intuitions that are directly about metaphysical matters, such as intuitions concerning causation or time, metaphysicians often find themselves relying on methodological intuitions as well. By "methodological intuitions", I mean intuitions about whether an argument is good, or an explanation is satisfactory, or in general intuitions about how to do metaphysical inquiry well. The nature and role of these intuitions raise very similar issues to questions about the nature and role of metaphysical intuitions, and similar debates can be had about their ultimate source.

Whatever the ultimate source and justification of intuitions, it would be difficult to engage in a lot of the development of theories and offering of counterexamples that goes on in philosophy without being able to rely on one's own sense of what seems correct and what seems like a mistake. Arguments can often be given for the positions one takes, but those arguments always need premises, and unargued-for premises seem unavoidable. Not anything goes in adopting a premise, however, and sometimes (usually?) we will get to a point where we will be better at saying what seems correct to us, metaphysically or methodologically, than providing a plausible argument for it. The practicing metaphysician at that point might rest, pro tem at least, with an appeal to "intuition"; and that may well be a respectable strategy whatever the ultimate story about intuitions themselves turns out to be.

4. Co-ordination with Science

Science, particularly natural science, seems to many to offer some of our best methods for finding out about the world. Understandably enough, many metaphysicians pay a lot of attention to what scientific theories have to say about metaphysical matters. Analytic metaphysicians are often accused of neglecting science, for example in chapter 1 of Ladyman et al. 2007. But I think this is just an error about the majority of contemporary analytic metaphysicians, who do keep at least one eye on the sciences. Metaphysicians are often wary of having their theories *conflict* with the best current science, and many want to go further, and use scientific findings to be a positive guide about which metaphysical theories are most likely to be correct.

One peculiarity of the current philosophical climate is that some philosophers who engage in metaphysical issues that interact with scientific ones primarily identify their activity as "metaphysics" while others identify it as "philosophy of science" (or perhaps as philosophy of some particular science: philosophy of physics or philosophy of

biology or philosophy of psychology, for example). If this reflected a deep methodological divide, then it might make sense to think that metaphysicians neglect science when developing metaphysical views. I am not sure it reflects any such thing, however: there is a fair amount of cross-citation between writers who self-identify in these different ways, and I suspect most writers on metaphysical topics in one camp are at most two or three steps of influence away from people in the other camp. In my view, both sides are doing work in analytic metaphysics, and so my remarks here about interaction with science should apply to both.

It is probably fair to say that physics gets more attention from contemporary metaphysicians than other parts of scientific inquiry, and within physics, fundamental physics gets the lion's share of attention. Perhaps this is because some topics of traditional metaphysical interest engage directly with topics of interest in fundamental physics: what are the smallest physical entities, for example, or what are the smallest divisions in space and time. Some of it might also be due to a philosophical view, explicit or implicit, about a hierarchy of natural scientific investigation with fundamental physics in a unique position of privilege (e.g. it might be thought that the ontology of fundamental physics is "fundamental" in some metaphysically charged sense). Whatever the sociological explanation, it should be kept in mind that other sciences, and for that matter other sub-fields of physics itself, have the potential to usefully inform metaphysics.

To take one recent example, Emma Tobin argues that there are examples in organic chemistry where it is plausible that determinate-determinable structures among natural kinds "crosscut": a determinate can belong to different kinds of determinables, neither of which is a sub-determinable of the other (Tobin 2010). To give one of Tobin's cases: renin is both a protein and an enzyme, but the kind *protein* neither belongs to, or includes, the kind *enzyme*. This goes against a traditional view of determinate-determinable structures, deriving from at least as far back as Aristotle, which is that at least when we are concerned with real natures of things, each determinate stands in only one structure of determinates and determinables.

As well as a connection at the level of content, with sciences casting light on metaphysical questions and perhaps metaphysics casting light on scientific ones, there are potential connections at the level of method as well. One approach to the method of metaphysics is to look at what methods are used in the sciences, and to see how well they apply to metaphysical questions. Some take this to show that metaphysics should be an a posteriori, naturalistic enterprise, eschewing the a priori. Others point to mathematics as an example of an apparently very successful enterprise, up to its neck both in questions of great generality and armchair, arguably a priori, methods. What the methods of the sciences are, and how they would be applied to traditional metaphysics questions, are both sufficiently controversial that there is little consensus about what metaphysics would look like if it used the methods of science (nor whether practicing metaphysicians *do* use the methods of science, more or less). My own view is that the methods of metaphysics are continuous with the methods of the deductive and natural sciences, and no doubt that opinion to some extent colours the discussion in this chapter. How to go from the general conviction that we can improve the method of metaphysics by considering the methods of other successful inquiries, to implementing practical improvements, is still a difficult and controversial issue.

5. Respect for Folk Opinion

The sciences are not the only source of empirical information about the world. Much of our information about the world comes from "common sense" unsystematized information gained through ordinary life. Over the course of their lives, even scientists can learn more from their parents than they do from their laboratories. Since information of this sort is the basis of a large part of our beliefs, it is not surprising that metaphysicians pay attention to it as well.

One way ordinary pre-scientific opinions play a role in metaphysical theorizing is as providing starting points for investigation. We might notice, for example, that matches can cause forest fires, but rain typically does not. This causes apparent trouble for theories of causation according to which, given other ordinary assumptions about the world, rain often does cause forest fires (see Mackie 1992). We might notice that sometimes unfortunate people lose feet, and that this gives rise to a puzzle about what happens to the rest of the person (see e.g. Olson 1997). We might notice that the sky is blue, which is a serious problem for positions in the metaphysics of colour where colour is a certain surface-reflection property. Sometimes our awareness of these commonplaces can seem too mundane to mention: but there is a skill in knowing which commonplace might play an important role in the development of a metaphysical theory, and it is worth keeping sight of the role of these commonplaces when someone claims that the source of metaphysical insight is all a priori, or all intuition, or a choice between the a priori and science, or similar sweeping methodological claims.

Respecting pre-theoretic opinions we gain through ordinary experience of the world may not be an entirely distinct practice from the previous three methods mentioned. Some of our ordinary beliefs might well be analytic or conceptual truths, and so be vindicated by conceptual or linguistic analysis. Some of our ordinary beliefs might have their origin in intuition (one candidate might be the ordinary belief that a whole is at least as large as any of its parts, for example). Some ordinary beliefs might be the result of past scientific investigation: it is an ordinary belief that the planets orbit the sun and together (with various moons, asteroids, dust, etc.) they make up the solar system: but this belief entered ordinary opinion from science. (The solar system is incidentally of metaphysical interest: if the solar system exists, it shows there can be wholes which have parts that are widely scattered.)

One way to downplay the role of ordinary beliefs in metaphysics is to maintain that whenever an ordinary belief was employed, a belief delivered through another, equally accessible, source would have been practically as good. Perhaps we need knowledge of the world to know that rain can delay a forest fire: but we could have detected the *possibility* of this through conceptual analysis or intuition or somesuch, and that this would have had the same impact on our philosophy of causation. I doubt this is always the case: sometimes the best way in practice to discover something is possible is to discover it is actual, and in general *possibly p* is not as strong a premise as *p*, and so does not always do the same work for us. But even if it is true that some resource available through conceptual analysis or intuition would be adequate for any given purpose in metaphysics, ordinary beliefs may still have a distinctive methodological role. They can be less controversial than claims about what is delivered by conceptual analysis or intuition—there may be a debate to be

had about whether intuition can tell us that non-surfaces could be blue, but less (or a different sort) about whether the sky is, indeed, sometimes blue, and whether common sense tells us so.

How beholden metaphysics is to ordinary beliefs is a controversial matter. Presumably not every widely held belief has to be respected: a metaphysics with no place for angels would not be thrown out just by discovering through opinion polls that many people believed in angels.[1] On the other hand, many would think that a metaphysics that claims there are no people, or no tables or chairs, or that nothing ever happened, would have its work cut out for it. There are live metaphysical views which disagree with common sense in just this way, but even those who hold such views often feel some pressure to justify this departure from ordinary opinion. One view to take is that a metaphysical theory does not have to agree with common belief (or some privileged core of common belief), but it must *accommodate* those beliefs: if it disagrees, it must explain how it entails something close to the ordinary beliefs, or how we can reinterpret what we ordinarily say or think, without too much violence to the ordinary opinions, so that the reinterpretation is compatible with the metaphysical theory in question.

It is also not an easy matter to determine when a metaphysical theory conflicts with an ordinary belief. A number of metaphysicians produce theories that on the face of it vary wildly from ordinary belief, but then adopt what Dan Korman calls "compatibilist strategies" to argue that when we properly interpret ordinary talk and belief the metaphysical views do not conflict with ordinary belief at all, or to no great extent (Korman 2015, section 7.1). (O'Leary-Hawthorne and Michael 1996 label this strategy as "offering a compatibilist semantics".) Peter van Inwagen, for example, claims that the only things with parts are living beings, and that there are no tables, chairs, mountains, planets, and so on. ("Tables are not defective objects or second-class citizens of the world; they are just not there at all" (van Inwagen 1990, pp. 99–100). However, he also claims that this "does not contradict our ordinary beliefs" (p. 98): what we ordinarily think and say when we say, for example, "there are two tables in my office" is true, despite the metaphysical facts. Whether the picture of language and thought van Inwagen is working with can be justified is of course controversial—but this is a stark example of a metaphysical thesis that is apparently in sharp disagreement with common-sense belief, but where things may not be so simple when the full details of the theory are examined.

One class of beliefs, usually associated with common-sense or ordinary opinion, which are sometimes thought to be particularly important for metaphysical method are so-called *Moorean beliefs*.[2] Some hold that there is a class of beliefs that are so certain that they are immune to philosophical revision. G. E. Moore cited his belief that he had a living human body as a belief so secure that he could reject philosophical theories which conflicted with that belief: the kinds of objective idealism, for example, which at least according to Moore denied the reality of material things (Moore 1925). (Furthermore, he thought his *knowledge*

[1] According to a 2007 Gallup Poll, 75% of Americans believe in angels. URL: <http://www.gallup.com/poll/27877/Americans-More-Likely-Believe-God-Than-Devil-Heaven-More-Than-Hell.aspx> (accessed October 10, 2015).

[2] Moorean facts, in something like the sense discussed, are mentioned a number of times by David Lewis: see, for example, Lewis 1983 pp. 351–7. Lewis derives the terminology from D. M. Armstrong: see Armstrong 1980, pp. 440–1.

that he had a body in such a case was so secure that he could thereby reject the sceptical hypothesis that nobody knows anything about the external world. Not only was it certain he had a body, but it was certain that he *knew* he had a body.)

If there is a class of *Moorean facts* in the world, truths corresponding to our Moorean beliefs, and if in addition we are often in a good position to tell what they are, presumably those truths should be constraints on our metaphysical theorizing. There could be Moorean facts in one of at least two ways. One would be if there were a domain of facts that had such a high degree of certainty so as to be almost incorrigible. (Perhaps Moore thought this.) The other would be if the epistemological credentials of metaphysics, and perhaps philosophy, were shaky enough that some beliefs were immune to threat from *philosophical* reasoning, even if they did not enjoy an exceptionally high degree of certainty. One could believe in Moorean facts either because of epistemic optimism about our capacity to detect those facts, or epistemic pessimism about philosophy's standing to challenge them.

6. Application of Theoretical Virtues

Once some constraints have been identified through other means, whether by intuition, deliverances of science, reflections on language or folk commitments, often there is still more work for a metaphysician to do. A metaphysician will often want to construct an explicit theory that respects the data, or respects most of the data (one reason not to incorporate it all is if some starting points conflict with others: if contrary intuitions are available, or common sense disagrees with science, or in some other way). Often, several theories can be constructed to fit with the data. What can a metaphysician do then?

One obvious thing to try is to collect more data—more intuitions about cases, more relevant scientific results, or whatever. Another thing to do is to employ some constraints in theorizing other than mere consistency with data. One theory might be *simpler* than another, for example. Simplicity seems to take several forms: a theory can be more parsimonious, postulating fewer kinds of entities or even fewer entities of each kind; it can have simplicity of formulation, where all of the content flows from a few concise, powerful axioms; and there are other aspects of simplicity as well. Some metaphysical theories seem to be more *explanatory* than others: one that postulates some underlying feature of a range of cases might seem more explanatory than one that is a hodgepodge of particular observations, for example. Sometimes a metaphysical theory will *unify* apparently disparate phenomena, as when many different kinds of ontology are specified from a few simple primitives (see e.g. Zalta 1983). Sometimes a theory will strike us as *beautiful* or *elegant*, and seem desirable on that count.

Several questions arise about the use of these sorts of considerations, which are often put together under the label "theoretical virtues". One is whether theoretical virtues like these are indicators of something epistemically valuable (truth, likelihood, justification, or somesuch), or merely pragmatically valuable (they make the theories easier to use, or give us pleasing aesthetic reactions, or make a theory easier to develop further, and so on). Of course, if these virtues are epistemically relevant they may be pragmatically relevant too, and it is relatively uncontroversial that, for example, simplicity might sometimes be

convenient. If these virtues are not epistemically relevant, however, then arguably they are relatively less important from the point of view of methodology. Many metaphysicians do think theoretical virtues like simplicity play an epistemological role, though here is not the place to try to sort out whether that is correct.

Another pressing question if we are interested in appealing to theoretical virtues to choose between theories is determining what the theoretical virtues are and their importance. There is no uncontroversial answer to these two questions: even determining when two names of virtues pick out the same thing can be tricky. Is unificatory power a species of simplicity, for example? Is elegance valuable as a mark of simplicity, or is it valuable in its own right? Are "fecundity" and "fertility" different labels for the same virtue, or names of slightly different phenomena? These questions can be difficult to answer, and given that usage in this area is not very settled, the best way for theorists to proceed may well be by stipulating what they mean by one of these expressions, when employing a consideration in theorizing.

Questions about the role and value of theoretical virtues are still matters of dispute, like much in methodology. Fortunately, we do not need a developed theory of theoretical virtues to be able to apply them in particular cases (just as a botanist can develop expectations about flowers on the basis of induction without needing a developed theory of inductive inference). Sometimes a metaphysical option will seem needlessly complex, for example, or a solution will strike a theorist as pleasing and elegant, or a development rejected as ad hoc, without needing or having a general theory of the range of conditions those criteria should be applied in, or what trade-offs are allowed or not. Such a theory might be desirable, and we risk misapplying virtues when we are not sure what they are and how they work, but there seems no reason to believe we would have to wait for that general theory before using our judgement in applying theoretical virtues in particular cases.

In this respect, construction of metaphysical theories may not be so different from construction of theories in other areas: in many areas of science, from physics to economics, theorists are supposed to do more than construct a theory consistent with available evidence. Features like providing simple explanations of data, unifying apparently disparate phenomena, offering theories which have applications beyond their initial birthplace, and avoiding ad hoc manoeuvres are all valued. Sometimes in the sciences some of the theory of how to use these considerations has been codified, for example in applying statistical methods that trade off simplicity and accuracy. But it is not as if scientific method is all codified and uncontroversial either: judgements about the theoretical virtues of theories often have to go well beyond what philosophy of science has managed to establish about theoretical virtues, let alone the theories of method consciously employed by individual working scientists.

7. Other Topics

Settling what resources analytic metaphysics has available to solve metaphysical problems does not settle all interesting methodological questions about this approach. While it is

impossible to list and discuss every methodological topic of interest that might arise, four other significant topics in the methodology of metaphysics are worth discussing here.

7.1 Formal Methods

One noticeable change in the way metaphysics was done in the twentieth and early twenty-first century, compared to earlier eras, is the widespread use of formal logic and some portions of mathematics, such as set theory, in metaphysical research and publications. On the face of it, this might seem surprising: metaphysics is for the most part a qualitative rather than quantitative inquiry, and the subject matter of metaphysics does not on the surface lend itself to mathematicization. Using formal methods in metaphysics is not an entirely novel development, of course: Aristotle employed syllogisms in the setting out of some of his metaphysical doctrines, for example, and so was consciously applying formal logic even if it was not symbolic logic in the style we have become familiar with since Frege and Russell.

Some metaphysical writing employs symbolic logic and similar formal machinery; some employs mathematical machinery, such a set-theoretic models or other uses of set theory; and some writing employs both. My view is that the explanation for using this apparatus will be different for different topics in metaphysics.

Sometimes a branch of metaphysics will inherit some of the use of formal machinery from its subject matter: someone engaged in providing a metaphysics of mathematics will likely employ some mathematical notation in describing the target of the theory, and will often find it useful to show, for example, that standard mathematical results still obtain in systems that are constructed. One example is when Lewis 1991 defends the view that the subset relation literally is a matter of part-to-whole, and that this sheds light on our understanding of the ontology of mathematics. One of the things he wants to do is to show that standard axioms of set theory all turn out to be correct in his system—and the need to show that these axioms are all indeed verified by the system he constructs involves a certain amount of formal logic and mathematics. Another example: if one is trying to provide a metaphysics for quantum mechanics, one would need to show that there are features of the metaphysical system proposed corresponding to a quantum-mechanical wavefunction and measurements, and to show that those features in the described system have the same mathematical characterization as they do in the standard quantum mechanical equations. Doing this would naturally lead to a certain amount of discussion of, and use of, the sort of mathematics found in quantum mechanics itself.

Another common way a metaphysical investigation can inherit formal machinery is when the investigation is closely connected to an investigation of language. Contemporary theories of syntax and semantics have become quite formal, and philosophers hoping to illuminate the truth-conditions of a puzzling piece of language may well reach for a toolkit with a lot of intensional logic and set-theory built in. (This raises the question of why, for example, contemporary semantics has become such a formal enterprise, but that is not a question I will attempt to answer here.) The extent to which metaphysicians ought to be interested in language is controversial: but whatever

the verdict on whether they *should* pay a lot of attention to language, while they do it will be hard to avoid complexities in our theories of how language works showing up in metaphysical discussions connected to that language. Another, more specific way the formal resources of a theory of language will have repercussions for metaphysics is when metaphysicians pursue questions concerning the metaphysics of language and meaning. An adequate theory of the nature of propositions, or truth-conditions, or meanings of component expressions (e.g. predicates, adverbs, sentential operators or quantifiers) will likely have a formal component if the linguistic theory to be relied upon itself is very formal.

Formal methods are often employed when the subject matter itself is not terribly formal. Whether there is a good sense in which I could have done otherwise, or whether God would be both inside or outside time, to take two examples, are not themselves questions that seem to require much formal logic or mathematics to ask. Still, a theory intending to respond to these questions might feature a fair amount of formalization. Theories of what could happen, or how an object could be different (e.g. how an agent could have acted otherwise) are likely to draw on modal logic, or an apparatus of possible worlds, or both. Theories of the relationship of God to time might naturally employ temporal logic, or a model of objects in times, or both. Sometimes questions that are not themselves very formal are related to formal questions: formal theories of modal and temporal operators, the part–whole relation, and some other matters, have been developed, and now those resources are available they are tempting to employ in other discussions of possibility and necessity, time, or parts and wholes.

Finally, formal methods, especially formal logic, are sometimes employed as part of an attempt to make an argument clear and rigorous. Many metaphysical discussions concern unfamiliar subject matters and difficult or paradoxical topics. Making sure that there has not been a slip in a series of arguments, and that the structure of the argument offered is communicated clearly, are both valuable. Going all the way to translating claims into formal logic is not always required: sometimes numbered premises rendered in "formalese" (e.g. "There is an object x, property P and time t such that …"), and an effort to make the arguments presented explicitly deductively valid, can often yield sufficient rigour without the need for mathematical or logical symbols. Metaphysics is of course not alone in having a need for explicit and rigorous arguments, and use of the tools of formal logic can be found across philosophy. (Arguably reliance on deductive argument in general is one of the core methods of contemporary metaphysics: it would have perhaps merited a section in the first half of this chapter except that it seems relatively unremarkable and uncontroversial.)

Of course, there is the danger of misplaced formal rigour—dressing up an argument in formal garb sometimes adds nothing to how informative or persuasive it is, and sometimes a predilection for lots of formal notation suggests the goal is more signalling that one is part of an in-group, or is the result of a habit of formalizing things with insufficient regard to whether formalism is useful for the particular purpose at hand. And formal notation can be worse than useless if it is badly defined or employed carelessly: often it is easier to see what someone might be getting at even if it not entirely clearly expressed when an argument is in English, but when formal notation becomes jumbled or inconsistently defined, the result can be complete garbage.

7.2 Tradition/Community/History

The traditions in which a metaphysician is writing can have an immense impact on his or her metaphysical assumptions, the questions she or he takes to be important, the rival positions engaged with in developing a conclusion, the style of approach to developing and defending positions; and other crucial aspects of a metaphysician's work. Despite this, discussion of how to critically reflect on the role these traditions play gets relatively little space in discussions of method in metaphysics. (My attempt to discuss this aspect of philosophical method at more length can be found in Nolan 2007.)

It is difficult to challenge all of one's starting points at once, and even if it seems to a philosopher that many of her or his starting points are "intuitive" or "common sense", which disputed premises a philosopher will be prepared to assert without argument often depends on which ones are taken for granted in the particular community she or he works in: many metaphysicians will assume that there are no non-existent objects, or that the law of non-contradiction holds exceptionlessly, though both positions are controversial and there have been times and places where those things could not have been assumed.

The effect of background and the community in which a metaphysician finds herself or himself working in can play a very large part in choices of topics to pursue in depth, approaches to those topics, and other aspects of a metaphysician's work other than just which opinions a metaphysician will start with in thinking, or premises the metaphysician will employ when arguing.

There is often a strong "family resemblance" between someone working in analytic metaphysics and his or her dissertation advisor: even when a student has apparently very different conclusions from an advisor, still the methods used to get there, a sense of the landscape of the debate, or a sense of which issues are important may still be shared. Some of that might be pre-selection (students have some choice over where they do their doctorates, and who at those institutions they do them with). Some of that is no doubt exposure to arguments—I expect most analytic metaphysicians flatter themselves they could convince another person of *some* part of their views if given a few years to do so. Some of it will result from guidance to work by others in the literature: given the size of the literature on most topics selective attention is necessary, and the papers and books a supervisor recommends are unlikely to be independent of the material that helped to shape that supervisor's own views. But, plausibly, some of this similarity is explained by less traditionally rational mechanisms—taking things for granted because they were taken for granted when one was learning about the topic, for example.

Human beings are social animals, and our tendency to pick up opinions, know-how and practices somewhat unconsciously from environments and communities is obviously not an entirely bad thing. Learning about a topic would be much slower if it could proceed only through being presented with explicit arguments to conclusions, for example. There does not seem to be a realistic alternative to being influenced by one's community and traditions when picking up unargued-for background assumptions, sets of issues to be tackled, or methods for proceeding in metaphysical argument. At least to some extent.

That said, there are some things one can do to minimize the risks that come with potentially uncritical absorption of these things from one's training and environment. One

obvious way that suggests itself is to familiarize oneself with other traditions and communities as well. These can be other strands of thought in metaphysics, work from people or journals outside one's usual stomping grounds, and sometimes it might be engaging people who are not primarily located in metaphysics: talking to colleagues who self-identify as philosophers of language or philosophers of science, for example. Most, or perhaps all, metaphysicians already make an effort to engage with people other than the "usual suspects", and this is among the benefits of that activity.

As well as engaging with contemporary traditions and communities, many metaphysicians spend a significant amount of their time engaged in studying the history of the discipline, whether it be the works of Hume, Aristotle, Kant, or other familiar names, or less widely read metaphysicians like Bosanquet or Duns Scotus or Proclus. The study of the history of metaphysics can be justified in its own right, as a significant and interesting part of the history of ideas. But many who engage in work on the history of metaphysics see it as having contemporary benefits. There is its function as a source of questions, theoretical options, and arguments for metaphysical conclusions, of course. But some of the benefits from engaging with the history of the discipline are more distinctive: seeing where one's own traditions come from can be useful for revealing why the philosophical landscape is the way it is today, and to make background assumptions explicit. Studying the history of the discipline also gives insight into very different ways of engaging with metaphysical topics, and in this way can be of benefit in the same sort of way that engaging with contemporaries beyond those in one's own tradition can.

One might reasonably ask whether metaphysics is distinctive, either insofar as different theorists belong to different lines of influence and take different matters for granted, or insofar as there is benefit to be gained from examination of its history. I expect there is at most a difference of degree between metaphysics and other areas of philosophy in this respect. However, there are some reasons to think that metaphysicians should pay particular attention to the variety of existing and historical approaches to metaphysical questions. One is that there is relatively little consensus about how to approach metaphysical questions, so it is harder for a single canonical set of issues or methods to be established. Another is that many of the questions of metaphysics are perennial: my sense is that it is easier for a metaphysician to find useful material in, for example Aristotle, that sheds light on issues of contemporary concern than, for example, a philosopher of language or a philosopher of contemporary science can. Metaphysics is unlikely to be the only sub-topic of philosophy as controversial in its methods and as connected with its history: ethics is arguably another such area.

7.3 What Epistemic Strength is Aimed At

Metaphysics speculates about some fairly abstract topics at times—the ultimate nature of reality, the metaphysics of unusual mathematical objects such as sets of inaccessible cardinality, the form of laws of nature, and so on. Metaphysics also addresses perennial topics: we have not yet been able to come to a consensus about whether general properties and relations exist, or whether there are any gods, or why there is something rather than nothing, despite debates about these topics (or relatives of these topics) stretching back

thousands of years. When a metaphysical theory concerns a distant subject matter, or an endlessly disputable one, then there is reason to worry about the epistemic status of even an apparently successful contemporary theory. Could it really have given us knowledge of its conclusions, after all these difficulties and all these disputes? And if that is not a reasonable expectation, then are metaphysicians being overconfident in making metaphysical assertions in the first place?

Different metaphysicians have different attitudes to what they are doing. Some are prepared to offer their theories as known, or at least good candidates for knowledge. Some only ask that the metaphysical theory they favour is better epistemically supported than rivals, even if that support is only weak (see Sider 2001, p. xv).[3] Others seem to be interested only in mapping out spaces of metaphysical options, and would claim no more for their favoured option than that it was interesting, or perhaps that it is the one they would like to be true. Many are likely to have different attitudes to different theories they propose: sometimes they have sufficient confidence in their conclusions to take them to be known; other times a position is described and defended in the cause of showing that it is another theoretical option to be taken seriously.

It is sometimes hard to tell what attitude a metaphysician takes to his or her own theory just from reading a presentation of it. Sometimes this is because a writer will just assert a view even though her own attitude towards it is more complex: often a writer will reasonably think a reader is more interested in the details of a position and arguments for it than the writer's own attitude to that view or arguments. Other times, articles will be filled with hedges, but it will not be clear how seriously they are to be taken. I have joked in the past that expressions like "it could be argued that … " or "a neglected alternative is … " function as the "British assertion sign", since many UK philosophers in particular have some tendency to present their own views together with such qualifiers. Sometimes there are good reasons for making qualified claims, even if the view to be presented is one the author fully believes. An author might not feel she has a sufficient argument for the view itself, for example.

Notice that even those who do not wish to claim they are in a position to *know* that their metaphysical doctrines are correct seem to have some significant epistemic commitments, ones which might already seem presumptuous to those sceptical of metaphysical inquiry. That position *X* is better supported than position *Y* will itself presumably be hard to establish if the common subject of *X* and *Y* is epistemically inaccessible. Even if a writer is only interested in mapping out the available options for dealing with a metaphysical problem, showing that one has mapped out all the options, or all the interesting options, is itself difficult. Even in areas that have received lots of attention, new theoretical options are often being uncovered and investigated, so if the long history of disagreement over a topic should make us suspicious that the latest paper contains the solution, it should probably also make us suspicious that the latest paper contains a correct list of the plausible potential solutions.

What could justify confidence in metaphysical conclusions, in the light of the apparent difficulty of the subject and the long track record of disagreement about metaphysical topics? One option is to think that justification is relatively local: if the particular argument

[3] For a recommendation that this should be our approach in philosophy generally, see Hájek 2007, pp. 229–30.

a metaphysician is considering *right now* is very plausibly valid and leads from very well-supported premises to a non-crazy conclusion, then that metaphysician might be justified in accepting the conclusion *even if* the field more generally is strewn with cases of metaphysicians apparently in that position with mistaken conclusions. (One might think seeing a barn lets one know it is a barn, even if one is surrounded by fake barn country.) Or we could have some reason to think that our contemporaries, or maybe only some of our contemporaries, are in a better general position to solve metaphysical problems: we have the latest scientific findings, the latest logical techniques, the latest surveys of the existing literature, or whatever. (Though we might be chastened by the thought that metaphysicians have thought that before and nevertheless made many errors.) Or perhaps there is no stable alternative than to take oneself to be largely right about the things one is confident about, and to work onwards from there. If that thought went along with the thought that one was automatically justified in doing so—which would require further argument—then the mere fact that the general epistemic situation might not look rosy ought not stop us.

Of course, this issue of the epistemic status that we should be aiming for theories to have is not a peculiar one to metaphysics. Many areas of philosophy are subject to extensive and long-term debate, and seem to concern topics that it would be difficult to know the truth about. Of course, this is not just a problem in philosophy: when we look at the track record of yesterday's orthodoxy being today's rejected error in the sciences, the "pessimistic meta-induction" is notoriously troubling. The trouble may seem even worse for metaphysics than for many areas of the sciences—at least in many mature sciences, a body of results that have stood the test of time can be gestured at. But there is no canonical list of accepted results in metaphysics that have stayed uncontroversial. (Perhaps this is only because such a list has never been brought to most metaphysicians' attention, but I doubt it.)

7.4 Metaphysics and the A Priori

One traditional conception of metaphysics is as an a priori discipline: indeed Kant thought that the central epistemological puzzle about metaphysics was the puzzle of how synthetic a priori knowledge was possible, since he thought that many of the central topics of metaphysics (e.g. the nature of causation, whether there are substances) were synthetic a priori. Many contemporary metaphysicians hold that some or all of their conclusions are a priori. On the other hand, there are a number of metaphysicians, especially those influenced by Quine, who are inclined to either reject the distinction between a priori and a posteriori, or at least to hold that many metaphysical conclusions are firmly on the a posteriori side of the distinction.

The extent to which metaphysics is a priori is not a question to be settled entirely separately from the other issues discussed in this chapter about appropriate method in metaphysics. If metaphysics was all a matter of linguistic or conceptual analysis, then presumably the results of metaphysics would have the same status as the results of those kinds of analyses—and the results of linguistic and conceptual analysis have traditionally been considered a priori. If an important role is played by intuition, then some metaphysical conclusions at least will inherit the status that intuitions have. Some theories of intuitions say that they are a priori, but others need not. If metaphysical conclusions draw on evidence

from the natural sciences, those conclusions are much more naturally considered a posteriori. Likewise for a lot of folk opinion—the doctrine that tables and chairs exist, or that the sky is blue, or that beliefs and desires cause actions, all seem to be learned through experience. The use of theoretical virtues in theory choice is a less clear case. I argue in (Nolan, 2015) that many of the uses of theoretical virtues should be seen as a posteriori as well, though this is controversial.

Some contemporary metaphysicians (myself included) find it strange that it should ever have seemed plausible that metaphysics was an a priori discipline. Why suppose we could tell general and deep facts about what the world as a whole is like without evidence from our senses? (It would be hard enough to discover such facts even with anything our senses could provide!) The impression that metaphysics is a priori may sometimes stem from a limited conception of what our senses could provide—it is no wonder that metaphysicians looked elsewhere than our senses if our senses could tell us no more than Locke or Hume thought they could. Another possible source of the impression that metaphysics is an a priori discipline is that the practice of metaphysics seems to be an "armchair" one. Metaphysicians do not have labs full of fancy ontology detectors, and they do not often go on fieldtrips to particularly inaccessible corners of Being. One explanation of this would be that the constraints on metaphysical theorizing are a priori and so available without any special sensory investigation of the world. But another is that the a posteriori evidence is easily available: this would be the case if ordinary understanding of the world was already enough to throw up significant philosophical problems, or if the scientific knowledge relied upon was already relatively well known among educated non-specialists. And, of course, if esoteric sensory discovery is sometimes needed for metaphysics, it could still be the case that most metaphysicians could employ it simply through reading those philosophers who have done the investigation, or read the specialist scientific reports of those who have.

One suggestion about why metaphysical inquiry had better be possible a priori is that ordinary judgements already implicitly presuppose substantial metaphysical assumptions. Kant's theory embodies one version of this (on one reading at least): the synthetic a priori truths about space and time are already presupposed in spatial and temporal judgements. Another recent version of this approach can be found in Lowe 1998. According to Lowe, ordinary a posteriori judgements often presuppose things about identity conditions of objects over time, or the nature of substances and properties, that must already be available, at least implicitly, which suggests the already available principles had better be available a priori. So some theorists do have reasons to insist that an important part of metaphysics must be available a priori, though these reasons normally go along with distinctive general epistemic positions.

8. Conclusion

As has been pointed out at various points, the methodology of metaphysics does not seem to be *sui generis*: techniques and challenges that arise in philosophical method appear elsewhere in philosophy, and indeed beyond philosophy. Is there, then, anything distinctive about the methods of metaphysics? My own view is that each of the topics discussed here

are relevant outside metaphysics as well, and, for example, Daly 2010 discusses a number of the same methods discussed here as methods for philosophy generally. If there is anything distinctive about metaphysics, methodologically speaking, it is more likely to be found in differences of degree rather than kind. Perhaps intuitions are relied upon more in metaphysics even than in other areas of philosophy. Perhaps respect for the metaphysical assumptions of common sense is stronger in metaphysics, for all the odd claims metaphysicians make, than in areas of inquiry such as chemistry. (Few articles in chemistry journals are concerned to reconcile recent advances with common-sense views of the interactions of different kinds of substances, for example.) Exactly what degrees of similarity and difference there are between metaphysics and other inquiries (inside or outside philosophy) is presumably a topic to be pursued through simultaneous investigation of method in metaphysics and method elsewhere. A detailed investigation of that question would thus fall outside the scope of this chapter.[4]

References

Armstrong, D. M. 1980. "Against 'Ostrich' Nominalism: A Reply to Michael Devitt". *Pacific Philosophical Quarterly* 61: 440–9.

Ayer, A. 1936. *Language, Truth and Logic*. Victor Gollancz, London.

Bealer, G. 1987. "The Philosophical Limits of Scientific Essentialism". *Philosophical Perspectives* 1: 289–365.

Bealer, G. 1996. "*A Priori* Knowledge and the Scope of Philosophy". *Philosophical Studies* 81: 121–42.

Black, M. 1952. "The Identity of Indiscernibles". *Mind* 61.242: 153–64.

Black, M. 1956. "Why Cannot an Effect Precede Its Cause?". *Analysis* 16.3: 49–58.

Daly, C. 2010. *An Introduction to Philosophical Methods*. Broadview, Buffalo.

Davidson, D. 1967. "The Logical Form of Action Sentences" in Rescher, N. (ed) *The Logic of Decision and Action*. University of Pittsburgh Press, Pittsburgh, pp. 37–71.

Hájek, A. 2007. "My Philosophical Position Says *p* and I Don't Believe *p*" in Green, M. and Williams, J. (eds) *Moorean Absurdity: Essays on Content, Context and Their Collision*. Oxford University Press, Oxford, pp. 217–31.

Jackson, F. C. 1998. *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford University Press, Oxford.

Jenkins, C. S. 2008. *Grounding Concepts*. Oxford University Press, Oxford.

Korman, D. Z. 2015. "Ordinary Objects". *The Stanford Encyclopedia of Philosophy*, Fall 2015 edition. URL: <http://plato.stanford.edu/archives/fall2015/entries/ordinary-objects/>.

Kripke, S. 1980. *Naming and Necessity*. Blackwell, Oxford.

Ladyman, J., Ross, D., Spurrett, D., and Collier, J. 2007. *Everything Must Go*. Oxford University Press, Oxford.

Lewis, D. 1983. "New Work For a Theory of Universals". *Australasian Journal of Philosophy* 61.4: 343–77.

[4] Thanks to Melissa Ebbers and Ole Koksvic for discussion, and especial thanks to Chris Daly, Dan Korman, Jonathan Schaffer, Jonathan Tallant, Jason Turner, and J. Robert G. Williams for comments on an earlier draft of this chapter.

Lewis, D. 1991. *Parts of Classes*. Blackwell, Oxford.
Lowe, E. J. 1998. *The Possibility of Metaphysics: Substance, Identity and Time*. Oxford University Press, Oxford.
Mackie, P. 1992. "Causing, Delaying and Hastening: Do Rains Cause Fires?". *Mind* 101.403: 483–500.
Moore, G. E. 1925. "A Defence of Common Sense" in Muirhead, J. H. (ed). *Contemporary British Philosophy* (2nd Series). Allen and Unwin, London, pp. 193–223.
Nolan, D. 2007. "Metaphysicians and Their Traditions". *Philosophical Topics* 35.1&2: 1–18.
Nolan, D. 2009. "Platitudes and Metaphysics". in Braddon-Mitchell, D. and Nola, R. *Conceptual Analysis and Philosophical Naturalism*. MIT Press, Cambridge MA.
Nolan, D. 2015. "The A Posteriori Armchair". *Australasian Journal of Philosophy* 93.2: 211–31.
O'Leary-Hawthorne, J. and Micheal, M. 1996. "Compatibilist Semantics in Metaphysics: A Case Study". *Australasian Journal of Philosophy* 74.1: 117–34.
Olson, E. 1997. "Dion's Foot". *Journal of Philosophy* 94: 260–5.
Sider, T. 2001. *Four-Dimensionalism: An Ontology of Persistence and Time*. Oxford University Press, Oxford.
Tobin, E. 2010. "Crosscutting Natural Kinds and the Hierarchy Thesis" in Beebee, H. and Sabbarton-Leary, N. *The Semantics and Metaphysics of Natural Kinds*. Routledge, London, pp. 179–91.
van Inwagen, P. 1990. *Material Beings*. Cornell University Press, Cornell Ithaca.
van Inwagen, P. 1997. "Materialism and the Psychological-Continuity Account of Personal Identity". *Philosophical Perspectives* 11: 305–19.
Wright, C. 1999. "Is Hume's Principle Analytic?". *Notre Dame Journal of Formal Logic* 40.1: 6–30.
Zalta, E. 1983. *Abstract Objects: An Introduction to Axiomatic Metaphysics*. D. Reidel, Dordrecht.

CHAPTER 10

PHENOMENOLOGY

TAYLOR CARMAN

1. Introduction

Phenomenology, though sometimes misleadingly referred to as a "method," or even a *discipline* in its own right, is better understood as a historical movement that arose at the beginning of the twentieth century, flourished for about fifty years, was overtaken by subsequent trends, and finally assumed its rightful place of honor alongside other such innovations as a style of inquiry and an enduring resource for philosophical thought.

The movement's founder, Edmund Husserl, and its three other major figures—Martin Heidegger, Jean-Paul Sartre, and Maurice Merleau-Ponty—were united by no explicit doctrine and in fact held widely different views concerning phenomenology's aims, limits, principles, and procedures. What they had in common, and so what gives the term its specific (though loose) methodological sense, was their shared commitment to thick, contextualized descriptions of lived experience and of the world as we encounter and understand it from a first-person point of view. Such description is and must be, they maintained, prior to conceptual abstraction, generalization, deductive argument, and explanatory theory. Husserl captured the spirit of his project in the slogan, "To the things (*Sachen*) themselves!" Of course, describing phenomena ("things themselves") is in principle an open-ended, potentially endless task. So, Heidegger writes in *Being and Time*, what is "essential" to phenomenology "does not lie in its being *actual* [or effective] (*wirklich*) as a philosophical "movement." Higher than actuality stands *possibility*. We can understand phenomenology only by seizing upon it as a possibility" (1962, p. 63).[1] The work of phenomenology is neither experiment nor demonstration, but rather, as Merleau-Ponty put it, "an infinite dialogue or meditation" (2012, p. lxxxv), which is why "phenomenology was a movement prior to having been a doctrine or a system" and "*allows itself to be practiced and recognized as a manner or as a style*," rather than a fixed method or a set of results (2012, p. lxxi).

[1] References to *Being and Time* are to page numbers of the German edition, given in the margins of the Macquarrie and Robinson translation, which I have modified throughout.

For his part, however, Husserl always intended phenomenology to be a robustly "scientific" (*wissenschaftlich*) enterprise, indeed a kind of *Ur*-science, grounding and legitimating all other sciences. But although favoring the concrete over the abstract, description over explanation, and qualitative over quantitative characterizations of its objects often put phenomenology at odds with the more positivistic and naturalistic currents of twentieth-century Anglo-American philosophy, the professional rivalry between phenomenologists and analytical philosophers tended to project a distorted image of the field. The two traditions often moved in different directions, but rarely found themselves in very clearly defined conflicts. So, just as the Vienna Circle and its latter-day critics and defectors developed tools that have facilitated work in logic and the philosophy of science and language, phenomenology has for its part inspired a century's worth of nuanced description and interpretation, especially in the arts, humanities, and social sciences.

Moreover, far from being inherently incompatible with other more abstract techniques and problem domains, phenomenology plays an indispensable role in all philosophical inquiry precisely by identifying and properly characterizing at the outset what arguments and theories must understand themselves to be arguments *about* and theories *of*. A theory of perception, for instance, must begin with some preliminary notion of what perception *is*. It is the business of phenomenology to ensure that such preliminary notions are coherent and plausible, that is, that theories will be grounded in and guided by adequate initial characterizations of their objects. Phenomenology has in this way made lasting critical contributions by exposing the distorted or unmotivated descriptive assumptions from which philosophical theories all too often proceed.

2. Husserl's Phenomenological Reductions

The word "phenomenology" first occurred in Johann Heinrich Lambert's 1764 *Neues Organon*, where it meant *doctrine of appearance*. What Lambert had in mind was not an autonomous science, but something more like a technique or corrective applicable to all inquiry, whereby subjective error and illusion could be eliminated in favor of appearances genuinely indicative of reality. In the early 1770s (indeed, in a letter to Lambert himself), Kant used the word to describe the project that would eventually become the *Critique of Pure Reason*, namely the prior demarcation of the limits of the cognitive faculties by which we know appearances, a demarcation requisite for the subsequent construction of any viable metaphysical system. For Kant, then, as for Lambert, "phenomenology" would be a corrective critical "discipline" only, not a positive source of knowledge (1997, A710/B738n). Hegel's *Phenomenology of Spirit* stands in this same lexical tradition, but with the added twist that, whereas "dialectic" had been for Kant a logic of speculative dead ends or *antinomies*, it now becomes in Hegel's phenomenology the way in which contradictions in the structure of appearance manage to be productive, generating and revealing rather than concealing the true essence of things.

Husserl's use of the word is very different. Although his kind of phenomenology can also be called a study of *appearance* as such, as opposed to the appearance of this or that particular thing, it is neither a tool for identifying and avoiding error in observation and

inference, nor a merely corrective discipline lacking its own positive results. It was instead to be a science of "pure" consciousness, that is, consciousness regarded not as an empirical psychological process, but as manifesting the *intentionality*—by which, following his mentor Franz Brentano, Husserl meant the *of*-ness or "aboutness"—of experience. The bulk of Husserl's philosophical work, from the turn of the century to the 1930s, thus consisted of (1) descriptions of the general structures and contents of intentional states of consciousness, especially perception, and (2) methodological accounts of how such phenomena can in principle be (or be rendered) objects of a *pure* (nonempirical) "rigorous science" of phenomenology (Husserl, 1981).

The two methodological innovations at the heart of Husserl's project are known as the *phenomenological reductions*. The most familiar of the two is the phenomenological *epochê,* or "transcendental" reduction, which involves "bracketing," or methodically setting aside from consideration, everything external to consciousness in favor of the internal contents of consciousness itself, contents Husserl believes are transparently accessible to the mind upon reflection.[2] There is, he insists, a "fundamentally essential difference," indeed a phenomenologically self-evident difference, "between *being as experience* and *being as thing*" (2014, p. 73–4). Here, he writes, "the intrinsic differentiation in the manners of being manifests itself, the most cardinal difference of all, that between *consciousness* and *reality*" (2014, p. 74). Consciousness and the world are radically distinct: "Between the senses of consciousness and reality, a veritable abyss yawns" (2014, p. 90). Transcendental subjectivity is, according to Husserl, fundamentally discontinuous with external nature: "Abysses separate everything [immanent to experience] from all nature and physics, and no less from all psychology—and even this image, as naturalistic, is not strong enough to indicate the difference" (2014, p. 177). The transcendental reduction, then, amounts to a special kind of reflection in which we concern ourselves not with the usual (outer) objects of our intentional attitudes, but with the (inner) contents of those attitudes themselves.

Taken by itself, however, the inward turn of the transcendental reduction would not be enough to distinguish phenomenology from introspective psychology. What is needed in addition is a turn from the empirical facts of consciousness to its "essential" structures, that is, from the material or stuff of experience to the intelligible forms that, Husserl argues, render our *subjective* experience *objectively* intentional, or world-directed. This reduction from content to form, or from fact to essence, is what Husserl calls the "eidetic" reduction (from the Greek *eidos*), an abstraction from contingent facts to the types or essences and the necessities they entail (2014, p. 5, *passim*).[3] The eidetic reduction turns our attention away from "real" mental events (occurring in time) toward "ideal" (abstract, atemporal) aspects of subjectivity that make possible its directedness to objects.[4] Such ideal aspects of experience include, for example,

[2] Husserl 2014, §§31–4, §§56–64. For other, slightly different accounts of the reductions, see also the fourth version of the article Husserl wrote for *Encyclopaedia Britannica,* and the reworking of that text in his 1928 Amsterdam Lectures on "Phenomenological Psychology" (both in Husserl 1997).

[3] By "essence" (*Wesen*) Husserl does not mean the *defining* property of a thing, but *any* property or form understood as ideal or general, as opposed to real or particular.

[4] Husserl's distinction between the empirically *real* and the *ideal* or "essential" aspects of mental phenomena in *Ideas* (1913; Husserl, 2014) had also been his chief concern in *Logical Investigations* (1900/01; Husserl, 2001), prior to his elaboration of the *epochê* and his emphasis on the "transcendental" status of subjectivity.

that ordinary sense perception intends objects in space with back sides, that all consciousness at a given time contains "protentions" of a future and "retentions" of a past, and that the selfsame intentional content can be in turn imagined in imagination, judged in judgment, and remembered in memory. Phenomenology is a science not of real psychological episodes and their contingent features, then, but of the ideal essences they necessarily instantiate. So, for Husserl,

> a phenomenological doctrine of essences is as little interested in the methods through which the phenomenologist could assure himself of the *existence* of the experiences (that serve as the underpinnings in his phenomenological determinations) as geometry is interested in how the existence of figures on the board or models on the shelf can be methodically secured. Geometry and phenomenology as sciences of pure essence [*Essenz*] make note of no determinations about real existence. (2014, p. 147)

3. From Husserlian Morphology to Heideggerian Hermeneutics

Heidegger was often at pains to distance his own conception of phenomenology from Husserl's. "'Phenomenology' in Husserl's sense," he once said, "was elaborated into a particular philosophical position already anticipated by Descartes, Kant, and Fichte. The historicity of thinking remained utterly foreign to it." According to Heidegger, that is, Husserl inherited from Cartesian rationalism and German idealism a conception of philosophy that clouded his view of the phenomena, contrary to his own stated aims. "The question of being developed in *Being and Time,*" Heidegger went on, "set itself against this philosophical position and on the basis of what I still today believe to be a more faithful adherence to the principle of phenomenology" (Richardson 1963, p. xiv).

What was the principle of phenomenology? In its most general and admittedly vague form, it is the exhortation to be descriptively true to "the things themselves." But what *are* those things? *Appearances*. The term "appearance," however, as Husserl observed, is ambiguous "between *appearing* and *that which appears*": you can have a *pleasant appearance* by looking nice, or you can *put in an appearance* by showing up. Husserlian phenomenology concerned itself with appearances in the former sense, not the latter: it was a science not of the objects of awareness, but of the contents of our awareness of them. So, although "*phainomenon* (phenomenon) in its proper sense means that which appears," in Husserl's technical vocabulary, "it is used primarily for the appearance itself, the subjective phenomenon" (1964, p. 11)—not things, but the *appearing* of things. Phenomenology was to be a science of subjectivity as such—a science of *seeming*, not of *being*.

But what about that very distinction between seeming and being, subjectivity and objectivity? Is *it* given phenomenologically? If so, where? How? Is it a self-evident fact, or perhaps an unexamined prejudice slanting Husserl's project from the outset? Like Nietzsche before him and Wittgenstein after, Heidegger rejected the metaphysical distinction between appearance and reality, since any notion of *seeming* presupposes some notion of *being*.[5] Appearances

[5] In the page-long section of *Twilight of the Idols* entitled "How the 'Real World' Finally Became

cannot be conceptually detached from our understanding of things actually showing up in some way, manifesting themselves—which is to say, *being* somehow: "appearance is possible only *on the basis of* something showing itself.... If one then says that with the word "appearance" we refer to something wherein something appears without itself being an appearance, the concept of phenomenon is not thereby defined, but *presupposed*" (1962, p. 29).

For Heidegger, then, a phenomenon is not an appearance or appearing in contrast to something that appears, but rather "that which shows itself, the manifest." Something showing or manifesting itself is a primitive notion that, he insists, "has in the first instance nothing whatever to do with what one calls 'appearance,' or indeed 'mere appearance'" (1962, pp. 28, 29).[6] Phenomena are not, as they were for Husserl, inner contents of conscious experience standing in representational or referential relations to the outer objects making their appearance in or through them. Phenomena are not *subjective*. The phrase "that which shows itself, the manifest" is what Heidegger in *Being and Time* calls a "formal indicator," a mere placeholder, which he then supplements with a more substantive "phenomenological" concept of phenomena. Phenomena in that more robust sense are not anything transparently present to conscious reflection, but rather hidden aspects of what lies open to view, something in need of evocation and interpretation. A phenomenon in the Heideggerian sense is

> something that first and foremost precisely does *not* show itself, something that, in contrast to what first and foremost shows itself, is *hidden,* but is at the same time something that essentially belongs to that which first and foremost shows itself, and belongs to it in such a way as to constitute its meaning and ground. (1962, p. 35)

The proper task of phenomenology, then, is not the kind of "pure" description urged by Husserl, but rather an attempt to draw out and highlight some obscure but important aspects of what shows itself. Moreover, fittingly, Heidegger construes the Greek *logos* as "letting something be seen" (1962, p. 33), so that phenomenology consists precisely in letting the ordinarily *unseen* dimensions of what is seen *be seen*. And what it calls for is not intuitive attention, but interpretive effort.

This is not to deny that phenomenology is an essentially descriptive project; indeed, Heidegger says the expression "descriptive phenomenology" is a tautology.[7] But whereas Husserl's work aspired to what he called "a systematic and eidetic morphology" of

a Fable," Nietzsche infers the incoherence of the metaphysical concept of appearance from the incoherence of metaphysical realism: "*with the real world we have also done away with the apparent one!*" (Nietzsche 1998, p. 20). In the same vein Wittgenstein writes, "To begin by teaching someone 'That looks red' makes no sense. For he must say that spontaneously once he has learnt what 'red' means.... Why doesn't one teach a child the language-game 'It looks red to me' from the first? Because it is not yet able to understand the rather refined distinction between seeming and being?" (Wittgenstein 1967, §418, §422).

[6] By "mere appearance" Heidegger (1962, p. 30) means the appearance of "what is essentially *never* manifest." What Kant calls a "phenomenon" is a mere appearance in this sense, since what appears through it is a thing in itself, or noumenon, which never shows itself as such.

[7] See also his remark a few pages later that the term "philosophy of life" (*Lebensphilosophie*) "says about as much as botany of plants" (Heidegger 1962, p. 46). The deeper substantive question, of course, is what *life* is, just as the deeper methodological question for phenomenology is precisely what *description* amounts to.

intentional attitudes (2014, p. 289), Heidegger disavows any purely observational notion of phenomenological description. Probably with Husserl's biological metaphor in mind, Heidegger writes, "Description here does not mean a procedure in the manner of, say, botanical morphology" (1962, p. 35). Rather, "the meaning of phenomenological description as a method is *interpretation*." In a word, "The phenomenology of Dasein is a *hermeneutic*" (1962, p. 37).

Heidegger's conception of phenomenology as essentially hermeneutical is therefore inconsistent with a stronger version of the principle of phenomenology, what the title of §24 of Husserl's *Ideas* deems "the principle of all principles," namely,

> *that each intuition affording [something] in an originary way is a legitimate source of knowledge, that whatever presents itself to us in "Intuition" in an originary way* (so to speak, in its actuality in person) *is to be taken simply as what it affords itself as, but only within the limitations in which it affords itself there*. (2014, p. 43)[8]

This principle of resting one's findings on primal, self-evident intuitions is a methodological application of a broader, systematic doctrine central to Husserl's phenomenology, both early and late. It appears in *Logical Investigations* as the concept of "categorial intuition"; in *Ideas* it becomes "essential intuition" (*Wesenserschauung*), or simply the "seeing" (*Sehen*) of essences. According to Husserl, "*Immediately 'seeing'—not merely sensory, experiential seeing* but *seeing in general, i.e., any kind of consciousness that affords [something] in an originary fashion*—is the ultimate source of legitimacy of all rational claims" (2014, p. 36).

4. Categorial Intuition and the Understanding of Being

It is widely assumed that Heidegger embraced Husserl's notion of phenomenological intuition, indeed took it for granted, with or without trying to render it consistent with his own hermeneutical approach. Heidegger would later recall how he was fascinated early on by the theory and practice of phenomenological seeing, particularly as Husserl describes it in the Sixth of the *Logical Investigations.* Husserl's early *magnum opus* seemed to emanate a "magic," a "spell" that captivated the young Heidegger, owing in part to its promise to offer insights into intentionality by carving a middle way between logic and psychology, between the purely formal and the empirically concrete (1972, pp. 74–8).

But Husserl felt he could justify his intuitive claims about intentionality only by showing intuition itself to be a legitimate source of philosophical evidence, capable of yielding valid generalizations, not just brute particulars. He therefore insisted that we have not just sensuous, but also "categorial," or logically structured, intuitions. We see objects (such as dogs) and their properties (such as black), but we also see—and not just in a metaphorical sense—that the dog *is* (or *is not*) black. Similarly, without having to see the pen and see the paper (in two distinct acts of seeing), we see the pen *and* the paper together on the desk: we literally

[8] Heidegger (1972, p. 63) refers explicitly to this passage.

see their conjunction or togetherness. So too, we see that *if* the glass falls, *then* the wine will spill onto the carpet, or that *either* the glass will *not* fall *or* the wine will spill, and so on.[9] According to Husserl, that is, we have concrete intuitions satisfying or fulfilling anticipations whose contents include formal elements such as *is* and *not,* the logical connectives *if, then, and,* and *or,* and quantifiers like *all, some, many, few, a,* and *none*. Moreover, contrary to the empiricists, we cannot derive logical or conceptual content from sensations simply through a process of abstraction. Instead, we must have insight into logically structured states of affairs that render our higher-order judgments true or false (Husserl 2001, II/§40).

This notion of a concrete acquaintance with the structure of intelligibility, in particular Husserl's claim that we have a direct intuition not just of entities, but of the phenomenal appearance of their *being* (and nonbeing), evidently made a deep impression on Heidegger. Understandably, he found himself drawn especially to §44 of the Sixth Investigation, where Husserl argues that we do not derive a notion of being from reflection on our own mental states, as if our understanding that something is (or is not) rested on an introspective observation of our own minds. Rather, more simply, we understand what is contained in the *is* precisely by being aware of states of affairs themselves. Not reflections on our experiences, but intuitions of things in the world are the source of our understanding that things *are*.

Husserl's theory of categorial intuition in *Logical Investigations* offered a powerful antidote to the intellectualism of neo-Kantian theories of mind, according to which all experience must be mediated by judgments and inferences. Indeed, Heidegger was drawn to Husserl's theory not for its assertion of the primacy of intuition, but for the claim that our fundamental understanding of being is something preconceptual, prior not just to reflection and introspection, but to conceptually articulated thought as such. This is why Heidegger would later remark that, for all its importance, Husserl's essential insight had already been "more primordially thought by Aristotle and in the whole of Greek thinking," which was of course also free of the Cartesian prejudices that had infected Husserl's work. Heidegger continues:

> The more decisively this insight became clear to me, the more pressing became the question, Whence and how is it determined what is to be experienced as "the thing itself" (*die Sache selbst*) in accordance with the principle of phenomenology? Is it consciousness and its objectivity or is it the being of entities in its unconcealedness and concealment? (1972, p. 79)

If Heidegger remained true to the spirit of the principle of phenomenology, then, it was only by repudiating the letter of Husserl's theory of categorial intuition, and with it Husserl's identification of "the things themselves" with the inner contents of subjective experience. Indeed, in *Being and Time*, his *magnum opus* of 1927, Heidegger argues that Husserl's emphasis on intuition is itself part of an ongoing Cartesian obsession with the mind understood as a self-contained region of being, abstracted from the shared world we actually inhabit:

> Under the unbroken hegemony of traditional ontology, the genuine mode of registering what truly is has been decided in advance. It lies in *noein*, "intuition" in the widest sense, from which *dianoein*, "thinking" (*Denken*), is simply derived as a founded form. And it is

[9] See Sartre's example (2003, pp. 34ff.) of *seeing* that Pierre is not here in the café at four o'clock, as opposed to judging idly that the Duke of Wellington and Paul Valéry are not.

> from this fundamental ontological orientation that Descartes gives his "critique" of the still possible intuitively apprehending mode of access to what is, *sensatio* (*aisthêsis*) as opposed to *intellectio*. (1962, p. 96)

Heidegger goes on to argue, moreover, that the restriction of intentionality to intuition and thought is part and parcel of the ontological prejudice he is most concerned to overturn, namely the assumption that all entities *are* in virtue of being substance-like or objective, hence ideally accessible to theoretical attitudes such as observation and judgment. Notwithstanding its laudable injunction to return to a concrete description of the phenomena, Husserl's phenomenology is, in Heidegger's eyes, just another case of the Western intellectual tradition's fixation on intuition, presence, and the temporal present:

> The thesis that all cognition has its goal in "intuition" has the temporal meaning that all cognition is a making present (*Gegenwärtigen*). Whether every science, or even philosophical thought, aims at a making present shall remain undecided here.—Husserl uses the expression "making present" (*Gegenwärtigen*) to characterize sense perception. . . . It was no doubt the *intentional* analysis of perception and intuition in general that suggested this "temporal" characterization of the phenomenon. The following division will show that and how the intentionality of "consciousness" is *grounded* in the ecstatic temporality of Dasein. (1962, p. 363n.)

In *Being and Time,* that is, Heidegger attempts to replace long-standing metaphysical and epistemological prejudices, including those still at work in Husserl's phenomenology, with a hermeneutic account of understanding as situated projection into future possibilities. Such an account is meant to expose the implausibility of the very idea that all human understanding rests on thoughts and intuitions directed to objects and objective states of affairs. The allusion to Husserl is unmistakable:

> By showing how all sight is grounded primarily in understanding . . . we have robbed pure intuition of its privilege, which corresponds noetically to the privileging of the occurrent in traditional ontology. "Intuition" (*Anschauung*) and "thought" are both derivatives of understanding, indeed rather remote ones. Even the phenomenological "intuition of essences" (*Wesensschau*) is grounded in existential understanding. (1962, p. 147)

Like his later notion of "essential intuition" (*Wesenserschauung*), Husserl's theory of categorial intuition has no place in Heidegger's phenomenology. Taken simply as a commitment to attend to concretely describable phenomena, the principle of phenomenology was, and remained, a source of inspiration for Heidegger; as a thesis of the primacy of intuition, it did not.[10]

5. Heidegger's Fundamental Ontology

According to Heidegger, philosophy is essentially *ontology*, the "science of being" (*not* of seeming) (1962, pp. 26, 230). Moreover, he maintains, "*Ontology is possible only as*

[10] For more on Heidegger's break with Husserlian phenomenology, see Carman (2003, pp. 53–100). Cerbone (2006) and Zahavi (2003) see more continuity between Husserl and the hermeneutical and existential phenomenologists, Heidegger, Sartre, and Merleau-Ponty.

phenomenology" (1962, p. 35). *Being and Time* thus brings together a striking combination of ancient and modern themes. The central question of the book concerns the meaning (*Sinn*) of being (*Sein*). What does it mean (for something) to be? What *is* it to be? The more immediate impetus behind Heidegger's phenomenology, however, was the questions he inherited from Husserl, namely, How does *intentionality* manifest itself in its most basic, essential, and concrete ways? How are these two kinds of inquiry related? What does intentionality have to do with the question of being?

What is *being*? The closest Heidegger comes to a definition is to say that being is that *in virtue of which* entities are entities. It is what *makes* (in a noncausal sense of "makes") entities, entities. More specifically, it is what we understand in our understanding of being, what we know when we know, however dimly and inarticulately, *what* and *that* entities are. Here, then, is our first clue to the link between the seemingly divergent ancient and modern themes in Heidegger's thought, the question of being and the problem of intentionality. For it turns out that to ask about the *meaning* of being is necessarily to ask about our *understanding* of being. Indeed, for Heidegger, being just *is* what we understand when we have an understanding of being. Being is entities *making sense* (to us) as entities—even if only tacitly, dimly, or altogether unconsciously. Unlike entities themselves, then, being manifests itself for, and so depends entirely upon, us.

Heidegger embarked upon two projects in relation to traditional ontology, only one of which he actually carried out in the two published divisions of *Being and Time*. He first sought to shed light on the structures of human existence and understanding that *make possible* all explicit accounts of *what* entities in general are. How was it ever possible for philosophers to pronounce on that subject? The answer, he suggests, lies in the fact that traditional ontology presupposes a more "fundamental ontology" rooted tacitly in the "pre-ontological" (that is, *pretheoretical, prereflective*) understanding of being that defines the entity we are, namely human beings (*Dasein*).

Being and Time is thus a kind of *transcendental* philosophy, that is, an account of the conditions of the possibility of something given, in this case our (sometimes explicit) understanding of *what* entities are and *that* they are. And just as Kant's transcendental critique of reason centered on an "analytic" (or *dissection*, Kant says) of the faculty of understanding, so too, Heidegger tells us, his own fundamental ontology—his explication of our average, everyday understanding of being—will take the form of an "analytic of Dasein," an interpretive interrogation of the entity we ourselves are. We shall ask ourselves, that is, how we experience and understand ourselves and the world around us, or more precisely how we experience and understand our *being* and the being of all the things we take *to be*.

So, for instance, whereas traditional ontology regarded all entities as objects, or substances with properties, Heidegger points out that ordinary things in our daily environment—tables and chairs, hammers and nails, doors and doorknobs, automobiles and street signs—do not show up in that way at all. A hammer is not a piece of steel adjoining a piece of wood, which we *first* see as an object with properties and *then* interpret or judge to be useful for hammering; it is first of all, and above all, something *to grab*, something *for hammering*. So too, a doorway is not just a rectangular aperture in a wall, but something *to walk through*. Such things are not occurrent or "present-at-hand" (*vorhanden*), but available or "ready-to-hand" (*zuhanden*). Even less is a human being a mere object with mental properties added on, but a doer and a sufferer, an agent

and a patient, not a *what* but a *who*, not some*thing* with extra psychological features in addition, but some*one* living a life, emerging from a history, and projecting into a future.

This is why Heidegger pursues the question of being by undertaking a phenomenology of human understanding. On the face of it, the two lines of inquiry might seem unconnected. Why take a detour through phenomenology, if our aim is to ask the question of the meaning of being? The answer is that being is just what manifests itself in an understanding of being, and since *our* understanding of being is the only one available to us, an interpretive description of *it*—in its concrete worldly setting—is the necessary path to even as much as grasping the question in its proper context.

6. Sartre and Merleau-Ponty on Consciousness and the Body

Sartre's 1943/2003 *Being and Nothingness: An Essay on Phenomenological Ontology* was, as both its title and subtitle suggest, inspired by *Being and Time*. Sartre, however, remained deeply attached to the Cartesian tradition in philosophy, and more particularly the broadly dualistic spirit of Husserl's account of consciousness as transcendental subjectivity.[11]

Adapting terminology from Hegel, Sartre draws a sharp metaphysical distinction between "being-in-itself" (*l'être-en-soi*) and "being-for-itself" (*l'être-pour-soi*), the latter referring to the reflexive structure of self-awareness exhibited principally, perhaps exclusively, in human consciousness. Human agents are beings-for-themselves; inanimate objects are beings-in-themselves. Alongside this ontological distinction, Sartre also describes what he calls "the double property of the human being," two distinct aspects of being-for-itself, namely "facticity" and "transcendence" (2003, p. 79). My facticity is that set of past and present facts about my body, my behavior, my character, and my social and physical situation, as they present themselves to a third-person point of view. My transcendence is the free, future-directed, first-person, conscious relation in which I stand to the world, including my own facticity. My facticity provides the setting and context of my transcendence, but my transcendence in turn determines what is salient and effective in my facticity.

Merleau-Ponty was deeply impressed and influenced by Sartre, but their ideas and their approaches to phenomenology differed profoundly.[12] Contrary to the sharp Cartesian distinction Sartre draws between the brute opacity of material reality and the, as it were, transparency and frictionlessness of consciousness, Merleau-Ponty regards human beings as *bodily subjects* embedded in essentially obscure and ambiguous ways in—hence neither mechanically determined by nor radically free in the face of—their natural and historical worlds. His *magnum opus* of 1945, *Phenomenology of Perception*, is a richly detailed

[11] In an interview in 1975, Sartre said, "I consider myself a Cartesian philosopher, at least in *Being and Nothingness*" (Schilpp, ed., 1981, p. 8).

[12] For more detailed discussion of Merleau-Ponty's conception of phenomenology and the relation of his work to Husserl's, see Carman (2008) and Romdenh-Romluc (2011).

description of the ways in which sense experience is instantiated in the body's semivoluntary and highly conditioned intuitive sense of its own spatial position, orientation, and repertoire of intelligent but noncognitive skills. His last, unfinished work, *The Visible and the Invisible*, extends and elaborates his earlier attempt to supplant the subject–object dichotomy with an account of what he calls the "interlacing" (*entrelacs*) of body and world (1968, ch. 4). Commenting on their very different styles and approaches to phenomenological reflection, Sartre later wrote, "Merleau-Ponty accepts thesis and antithesis. It is synthesis which he rejects … Spirals … are never allowed to conclude" (1965, p. 212).[13] In contrast to Sartre's regularly dramatic—often blunt, counterintuitive—conclusions, that is, Merleau-Ponty was attuned to the complexity, the ambiguity, the unclassifiable richness of the phenomena he sought to describe.

In spite of his debt to Husserl, Merleau-Ponty's phenomenology poses a radical challenge to all forms of *representationalism*, including Husserl's, which conceives of consciousness as an inner domain of intelligible content, sharply delineated from the outer world to which it purports to refer. Merleau-Ponty's phenomenology, by contrast, emphasizes the practical and bodily immersion of perception in the perceived world. Consciousness cannot be abstracted from the world, he maintains, since it is an aspect of human existence, which is essentially, as Heidegger says, "being in the world" (*être au monde*).

The Berlin school of Gestalt psychology was also a major influence on Merleau-Ponty's work. Like phenomenology, Gestalt theory emphasized the nonconceptual, prelogical coherence of perceptual experience. Perception is constituted by neither spontaneous judgment nor the mere passive registration of sense data, both of which presuppose the very perceptual coherence they purport to explain. What Merleau-Ponty took from the Gestalt school more particularly was its critique of the *constancy hypothesis*, namely the assumption of a one-to-one correspondence between sensory stimulus and perceptual content. The constancy hypothesis is the deep error common to both *empiricism* and *intellectualism*, the two then dominant positions in the philosophy and psychology of perception: according to empiricism, perception is grounded in sensation; according to intellectualism, it is grounded in judgment. A sensation is supposed to be the discrete effect of a sensory stimulus, yet what we experience in perception is not a fleeting mosaic of sensations, as empiricism suggests, but a stable and coherent *world*. Intellectualism, by contrast, recognizes the intelligibility of the perceived world and acknowledges that perception is not just a brute confrontation with sense data, yet it too takes the constancy hypothesis for granted precisely by inferring that perceptual content cannot come from the supposedly brute data of sense experience, but must instead be supplied by a nonsensory faculty, namely judgment.

Merleau-Ponty argues that neither the empiricist nor the intellectual approach to perception can be correct. To begin with, perception cannot be grounded in sensation, since the very concept of sensation is confused. What we see, after all, are not sensations, but full-fledged *things*. Indeed, words supposedly descriptive of mere sensations ("burning," "ringing," "spot," "patch") are in fact abstract, fragmentary bits of language originally referring to features of objects (fires, bells, swatches of cloth). The very idea of sensation is parasitic

[13] In the same vein, Sartre writes, "Alone, each of us was too easily persuaded of having understood the idea of phenomenology. Together, we were, for each other, the incarnation of its ambiguity. Each of us viewed the work being done by the other as an unexpected, and sometimes hostile deviation from his own" (1965, p. 159).

on our prior understanding of objects. Worse, it is caught between two often incompatible tasks: to describe the immediate *stimulus* to our senses, and to describe how things *seem*. The perceptual appearance of things differs widely from the array of sensory stimuli, though, so the concept of sensation can have no consistent role to play in a theory of perception. Neither can perception be grounded in acts of judgment, which presuppose something given, *about which* one can judge. By insisting that perception is judgment "all the way down," intellectualism robs itself of the very phenomenon that gave it content, namely judgment understood as an attitude taken up with respect to a world already somehow *given* in perception.

What both empiricism and intellectualism lose sight of, Merleau-Ponty maintains, is the *phenomenal field* itself, that is, the givenness of the world to a situated bodily perspective that is neither merely sensory nor intellectual in character. The *unity* of perceptual objects as such, problematized in the seventeenth century by Molyneux's famous question concerning the relation between visual and tactile sensory content, is grounded in the unity of the human body, "the material subject of the world" (Todes, 2001). So, for example, I see the flickering intangible form, the shining orange color, the searing heat, the burning odor, and the crackling sound all as constituting a single thing—*fire*—because I see it with my whole body, not just my eyes.

In making the *ontological* claim that perception is not a mental but a *bodily* phenomenon, Merleau-Ponty is neither merely reporting a subjective impression nor advancing a constructive theory about the objective nature of reality. He is instead describing, articulating, and clarifying the ordinary intuitive point of view from which we understand ourselves as neither disembodied intellects nor physical mechanisms, but living bodily subjects. That phenomenological or bodily perspective is what both traditional and contemporary theories of perception fail to recognize as a proper subject of inquiry, let alone the proper framework within which to understand intentionality. The body functions as the subject of perceptual experience by means of what Merleau-Ponty calls the *body schema*, the set of abiding noncognitive dispositions and capacities that orient, guide, and inform our bodily sensitivities and motor actions. To say that perception is grounded in the body is to say that the phenomenal field is constituted by the body schema. Our bodily skills and dispositions carve out a perceptual world with perspectival horizons and a contrast between figure and ground.

The kind of intentionality made possible by the body schema is not mental but what Merleau-Ponty calls "motor intentionality" (2012, pp. 113, 523ff n.99). Brain-damaged patients who suffer from visual form agnosia retain many motor skills and are able to think abstractly about spatial relations, but have lost an intermediate intuitive *motor intentional* sense of spatial relations, including their own bodily position and orientation (Milner and Goodale, 1996). Space is no longer *given* in their intuitive awareness, but now resides in pathologically segregated domains of "blind" motor action on the one hand, and decontextualized judgments about spatial relations on the other.

7. Legacy and Relevance

Contrary to Husserl's ambition and hope, phenomenology never proved to be a productive science, an autonomous discipline, an organized school, or even a well-defined method or

technique. Instead, it was a movement, a style, a discourse, and in the end an important chapter in the history of modern thought. Like all enduring contributions to philosophy, it continues to resonate in contemporary work across a wide range of disciplines, explicitly or implicitly, even when those who feel its influence officially declare themselves against it.[14] A recent collection of essays surveying phenomenology and its legacy includes chapters not only on its central figures, key concepts, and leading themes, but also on its impact on virtually every major subfield in philosophy, as well as deconstruction, feminism, critical theory, race theory, cognitive science, psychoanalysis, literary theory, and nursing and medicine (Luft and Overgaard, 2012). Another volume of scholarly papers purports to bring the phenomenological tradition up to date by linking it to recent developments in cognitive science, thereby, as the book's title declares, "naturalizing phenomenology" (Petitot et al., 1999). That would be anathema to Husserl and Heidegger, though perhaps not to Merleau-Ponty, who, alone among the major figures in the tradition, sought to forge direct links between phenomenology and developmental psychology, linguistics, anthropology, and biology.[15]

References

Carman, T., 2003. *Heidegger's Analytic: Interpretation, Discourse, and Authenticity in "Being and Time."* Cambridge: Cambridge University Press.

Carman, T., 2008. *Merleau-Ponty*. London: Routledge.

Cerbone, D. R., 2006. *Understanding Phenomenology*. Chesham: Acumen.

Foucault, M., 1970. *The Order of Things: An Archaeology of the Human Sciences*. New York: Random House.

Hegel, G. W. F., 1977. *Phenomenology of Spirit*. A. V. Miller, trans. Oxford: Oxford University Press.

Heidegger, M., 1962. *Being and Time*. J. Macquarrie and E. Robinson, trans. New York: Harper & Row.

Heidegger, M., 1972. *On Time and Being*. J. Stambaugh, trans. New York: Harper & Row.

Husserl, E., 1964. *The Idea of Phenomenology*. W. P. Alston, trans. The Hague: Martinus Nijhoff.

Husserl, E., 1981. "Philosophy as Rigorous Science." Q. Lauer, trans. *Husserl: Shorter Works*. P. McCormick and F. Elliston, eds. Notre Dame: University of Notre Dame Press.

Husserl, E., 1997. *Psychological and Transcendental Phenomenology and the Confrontation with Heidegger (1927–1931)*. T. Sheehan and R. E. Palmer, eds. and trans. Dordrecht: Kluwer.

Husserl, E., 2001. *Logical Investigations*. 2 vols. J. N. Findlay, trans. London: Routledge.

Husserl, E., 2014. *Ideas for a Pure Phenomenological Philosophy. First Book: General Introduction to Pure Phenomenology*. D. A. Dahlstrohm, trans. Indianapolis: Hackett.

Kant, I., 1997. *Critique of Pure Reason*. P. Guyer and A. W. Wood, trans. Cambridge: Cambridge University Press.

[14] Foucault, for example, in his Foreword to the English edition of *The Order of Things* (1970, p. xiv). My point is not that Foucault was doing phenomenology in that book in spite of himself, but that his project—unlike, say, the structural anthropology of Lévi-Strauss—was possible only on the basis of a deep familiarity and engagement with the work of the phenomenologists.

[15] See especially Merleau-Ponty's lectures from the late 1950s (Merleau-Ponty, 2003).

Luft, S. and Overgaard, S. eds., 2012. *The Routledge Companion to Phenomenology*. London: Routlege.

Merleau-Ponty, M., 1968. *The Visible and the Invisible*. A. Lingis, trans. Evanston: Northwestern University Press.

Merleau-Ponty, M., 2003. *Nature: Course Notes from the Collège de France*. R. Vallier, trans. Evanston: Northwestern University Press.

Merleau-Ponty, M., 2012. *Phenomenology of Perception*. D. A. Landes, trans. London: Routledge.

Milner, D. A. and Goodale, M. A., 1996. *The Visual Brain in Action*. Oxford: Oxford University Press.

Nietzsche, F., 1998. *Twilight of the Idols*. D. Large, trans. Oxford: Oxford University Press.

Petitot, J., Varela, F., Pachoud, B., and Roy, J.-M. eds., 1999. *Naturalizing Phenomenology: Issues in Contemporary Phenomenology and Cognitive Science*. Stanford: Stanford University Press.

Richardson, W. J., 1963. *Heidegger: Through Phenomenology to Thought*. The Hague: Martinus Nijhoff.

Romdehn-Romluc, K., 2011. *Merleau-Ponty and "Phenomenology of Percpetion*." London: Routledge.

Sartre, J.-P., 1965. *Situations*. B. Eisler, trans. Greenwich: Fawcett Publications.

Sartre, J.-P., 2003. *Being and Nothingness: An Essay on Phenomenological Ontology*. H. Barnes, trans. London: Routledge.

Schilpp, P. A. ed., 1981. *The Philosophy of Jean-Paul Sartre*. La Salle: Open Court.

Todes, S., 2001. *Body and World*. Cambridge, MA: MIT Press.

Wittgenstein, L., 1967. *Zettel*. G. E. M. Anscombe and G. H. von Wright, eds. Oxford: Blackwell.

Zahavi, D., 2003. *Husserl's Phenomenology*. Stanford: Stanford University Press.

CHAPTER 11

THE PRAGMATIC METHOD

HENRY JACKMAN

1. PRAGMATISM AND THE "PRAGMATIC MAXIM"

While classical pragmatism quickly became identified with the theory of truth that dominated critical discussions of it, both of its founders, Charles Sanders Peirce and William James, understood pragmatism essentially as a *method*. (The so-called, "pragmatic theory of truth" was originally intended to just be an instance of that method's application, albeit a very important one).[1] This pragmatic method plays a central role in Peirce's 1878 paper, "How to Make Our Ideas Clear", but it was introduced to the wider philosophical community (and first introduced by that name) by James, who in an 1898 lecture at the University of California at Berkeley,[2] presented the "principle of pragmatism" as the view that:

> To attain perfect clearness in our thoughts of an object, then, we need only consider what effects of a conceivably practical kind the object may involve—what sensations we are to expect from it, and what reactions we must prepare. Our conception of these effects, then, is for us the whole of our conception of the object, so far as that conception has positive significance at all.[3]

[1] When James defined pragmatism for *Baldwin's Dictionary* in 1902 (James 1902b, p. 94), there was no mention of any theory of truth, but by the time *Pragmatism* was published in 1907, it was becoming clear that the name "pragmatism" was being used a label, not only for the pragmatic method, but also for the particular account of truth associated with Peirce, James, Dewey, and Schiller. James initially preferred keeping the term "pragmatism" for the method and Schiller's term "humanism" for the theory of truth (see James's 1904 letter to Schiller (in James 1907, p. 163), as well as his 1904 article "Humanism and Truth", reprinted in James 1909 (especially pages 37–38)). However, by 1907 he seemed resigned to the fact that "pragmatism" was being used for both, so while he insisted that pragmatism "is a method only" (James 1907, p. 31), he admits that "the word "pragmatism" has come be used in a still wider sense, as meaning also a certain *theory of truth*" (James 1907, pp. 32–3).

[2] Published that same year by the University Chronicle as "Philosophical Conceptions and Practical Results" (James 1898). Since the Chronicle was only circulated in the University of California system (aside from the 30 reprints that James received), James published it again (in a slightly revised form) in *The Journal of Philosophy* under the title "The Pragmatic Method" (James 1904).

[3] James 1898, p. 259, James 1904, p. 124, 1907, p. 29. Virtually the same passage appears in *The Varieties of Religious Experience* (James 1902a, p. 351), and for most of James's contemporaries, that version would be their first exposure to the maxim.

James focused primarily on how the principle (hereafter "the Pragmatic Maxim") could be applied to solving disputes, particularly philosophical ones. He famously re-introduces the maxim in his *Pragmatism* by describing his application of it on a camping trip to a debate between two groups of his friends about whether a man would be "going around" a squirrel that kept itself on the opposite side of tree that the man was circling. James suggested that the answer "depends on what you *practically mean* by 'going round' the squirrel" (1907, p. 27). If you mean by "go round" being to the north, then to the west, then to the south, etc., then the man would have, while if you meant facing the squirrel's front, then facing its side, then facing its back, etc., then the man would not have. Both groups predicted the same experiences from their seemingly opposing claims, so James concluded that their debate was idle. Of course, resolving one's friends' campground disputes is a fine thing to do, but James was most interested in the maxim's application to "philosophical disputation", where he expected it "wonderfully to smooth out misunderstandings and to bring peace" and to "yield a sovereignly valuable rule of method for discussion".[4] As he puts it in *Pragmatism*:

> The pragmatic method is primarily a method of settling metaphysical disputes that otherwise might be interminable. Is the world one or many?—fated or free?—material or spiritual?—here are notions either of which may or may not hold good of the world; and disputes over such notions are unending. The pragmatic method in such cases is to try to interpret each notion by tracing its respective practical consequences. What difference would it practically make to anyone if this notion rather than that notion were true? If no practical difference whatever can be traced, then the alternatives mean practically the same thing, and all dispute is idle. Whenever a dispute is serious, we ought to be able to show some practical difference that must follow from one side or the other's being right.[5]

James goes on to claim that:

> It is astonishing to see how many philosophical disputes collapse into insignificance the moment you subject them to this simple test of tracing a concrete consequence. There can *be* no difference anywhere that doesn't *make* a difference elsewhere—no difference in abstract truth that doesn't express itself in a difference in concrete fact and in conduct consequent upon that fact, imposed on somebody, somehow, somewhere and somewhen. The whole function of philosophy ought to be to find out what definite difference it will make to you and me, at definite instants of our life, if this world-formula or that world-formula be the true one.[6]

James applies this method to the questions of whether the world has a material or spiritual origin,[7] and whether reality is ultimately "one" or "many",[8] and to philosophical problems relating to substance,[9] absolute idealism,[10] free will,[11] possibility,[12] intentionality,[13] and, most famously, truth.[14] In each of these cases, he hoped to support his thesis that in "every genuine metaphysical debate, some practical issue, however conjectural or remote, is involved".[15]

[4] James 1898, p. 259, (also in James 1904, p. 124). [5] James, 1907, p. 28.
[6] James 1907, p. 30. (All but the last sentence also in James 1898, p. 260, 1904, pp. 124–5).
[7] James 1898, p. 260, 1904, p. 125, 1907, p. 50. [8] James 1907, p. 63.
[9] James 1907, p. 45. [10] James 1907, p. 41. [11] James 1907, p. 60.
[12] James 1907, p. 136. [13] James 1909, p. 68.
[14] James 1907, ch. 6. [15] James 1907, p. 52.

Unfortunately, the application of James's method to philosophical cases proved far less successful than its application to his friends' campground dispute. Indeed, there remains a surprising amount of disagreement about just what the maxim James proposes actually is. In particular, there has been a considerable lack of clarity about just what sort of "significance" James is talking about, and most prominently, just what he means by "practical consequences".[16] There have been a number of interpretations of the Maxim, of which four of the most central are: the "Peircian" reading, the "activist" reading, the "subjectivist" reading, and finally the "practical" reading. While some of these readings have advantages over others (both in terms of exegetical plausibility and fruitfulness of the resulting maxim), none have proved entirely satisfactory, and so it is not surprising that contemporary philosophers inclined to describe themselves as pragmatists do so for reasons other than a commitment to any version of the Pragmatic Maxim.

2. The "Peircian" Reading

A natural reading of James's Pragmatic Maxim would be to take it to be essentially the one that Peirce introduced in 1878. James certainly encourages this when he introduces the maxim not only as "the principle of pragmatism" but at the same time as "the principle of Peirce",[17] and James's maxim does seem remarkably similar to Peirce's now familiar claim that in order to attain the third grade of clearness[18] with our ideas we:

> Consider what effects, that might conceivably have practical bearings, we conceive the object of our conception to have. Then our conception of these effects is the whole of our conception of the object.[19]

On the Peircian reading of the Pragmatic Maxim, the "significance" involved is *cognitive* significance, and the cognitive content of a philosophical view is to be understood in terms of the "practical effects" (in particular, the experiences) that would follow from its *truth*.

It was certainly this reading that led many to view the Pragmatic Maxim as being a variant of the 'verificationist' accounts of meaning that were beginning to gain traction in the early part of the twentieth century,[20] and it is unsurprising that the two were often viewed

[16] Indeed, such differences are often understood as explaining the difference between (in the terminology of Rescher 2000, pp. 64–5) the "pragmatism of the right" associated with writers like Peirce, C. I. Lewis, Haack, and Rescher himself, and the "pragmatism of the left" associated with James, Schiller, and Rorty. For a similar narrative seeing these two strains coming out from the interpretation of the pragmatic maxim, see also Mounce 1997 and Misak 2013.

[17] The maxim is so described in James 1898, p. 259, 1902a, p. 351, 1904, p. 124, and 1907, p. 29.

[18] The first two grades coming from familiarity and explicit definitions (Peirce 1878, p. 126).

[19] Peirce 1878, p. 132. This "method of ascertaining the meaning of hard words and of abstract concepts" is later identified by Peirce as "pragmatism", and, like James, Peirce prefers to think of pragmatism as a method only, rather than a "doctrine of metaphysics" or "attempt to determine any truth of things" (Peirce 1907, p. 400). Peirce, of course, modified and clarified his presentation of the maxim in the years that followed (for a discussion of this, see Hookway 2004, Misak 2013).

[20] See Soames, chapter 3 of this volume.

as notional variants of the same basic view.[21] Indeed, provided that one understood "practical consequences" to simply be "perceptual consequences", it would be hard to see how the two maxims differed. Furthermore, in Peirce's own hands, pragmatism had a noticeable "anti-metaphysical" tone. While James emphasized how the pragmatic method would allow us to resolve many philosophical disputes, Peirce at times seemed more inclined to treat pragmatism as a method to separate the problems that were intellectually tractable from the ones that were, ultimately, nonsense.[22] When he asked himself what he expected of pragmatism, Peirce provided an answer that would not have seemed out of place in the Vienna Circle:

> It will serve to show that almost every proposition of ontological metaphysics is either meaningless gibberish . . . or else downright absurd; so that all such rubbish being swept away, what will remain of philosophy will be a series of problems capable of investigation by the observational methods of the true sciences.
>
> (Peirce 1905, p. 338)

While the more expansive notion of experience that Peirce endorses allows him to take a different line on the meaning of mathematical sentences than the positivists,[23] the Pragmatic Maxim, understood in this Peircian fashion, faces a number of challenges similar to those faced by defenders of verificationist theories of meaning.

One such problem is that by tying the cognitive content of a statement to what practical effects we should expect if it were true, the Peircian interpretation of the maxim seems to leave without significance, statements about the past that don't have predicted consequences for the future. If I claim that Plato stubbed his toe 18 days before his eleventh birthday, the claim certainly seems meaningful, but it seems unlikely that there are any *particular* experiences that we can expect from it. Of course, Peirce insists that "it is unphilosophical to suppose that, with regard to any given question (which has any clear meaning), investigation would not bring forth a solution of it, were it carried far enough",[24] but most of us would be more skeptical about whether such cases would bring forth any such solution. Further, even in those cases where we have such predicted effects, the maxim still suggests that such statements are, in some sense, really *about* the future. If I claim that a large meteor landed 10,000 years ago where Toronto is now located, there may be future experiences (about, say, how an excavation would turn out) that this would predict, but to say that those predictions are the entire

[21] Ayer, for instance, claims that "Peirce's pragmatic maxim is indeed identical, for all practical purposes, with the physicalist interpretation of the verification principle" (Ayer 1968, p. 45). See also Lewis 1934, p. 65, Carnap 1936, p. 123, and Reichenbach 1938, p. 49. (For an account stressing the differences between the Peircian and the verificationist maxims, see Misak 1995.)

[22] However, see Peirce 1905 (p. 339) on how, unlike other "prope-positivist" theories, his maxim allows one to extract the "precious essence" of metaphysics (particularly as it relates to the nature of signs and categorical schemes). For a discussion of this, see Nagl 2004, p. 13, Haack 2006, p. 145.

[23] There is good reason to think that Peirce had much more than sensations in mind when he spoke of experience. In particular, he speaks not only of "external" experiences (of which sensation would be paradigmatic) but also of "internal" experiences, which would include, for instances, the experiences we have when engaged in a mathematical proof or the manipulation of a geometrical diagram. (For a discussion of this, see Misak 2013, p. 42.)

[24] Peirce 1878, p. 140.

significance[25] of my statement seems to leave out the fact that it is essentially a claim about what happened 10,000 years ago, not about what would happen if you took soil samples in Toronto tomorrow.

James seems willing to bite such bullets when he argues (in an example that will come up again) that, if there were no future, the dispute between those who think that God created the world and those who think that it resulted from "blind physical forces"[26] would (if both theories could explain all of our current experience)[27] be "purely verbal", and that "the two theories, in spite of their different-sounding names, mean exactly the same thing".[28] It is only because we have a future that such a debate between the theist and materialist is a significant one. The materialist is, according to James, committed to all life in the universe eventually perishing (and all of our accomplishments and values going unremembered), while the theist predicts that our values will be preserved even if we, as individuals, die out. As James puts it, "Materialism means simply the denial that the moral order is eternal, and the cutting off of ultimate hopes; spiritualism means the affirmation of an eternal moral order and the letting loose of hope."[29]

Another problem for the "Peircian" version of the Pragmatic Maxim is that it may seem that normative statements about they way things ought to be (rather than how they are) will not come out as meaningful when the maxim is applied to them. Such statements seem to have no predictive import, which would suggest that no practical consequences follow from their truth. There may be, for instance, nothing we can predict from the truth of a statement like "eating meat is wrong", since we may very well continue to eat meat forever in spite of its being wrong. Their denial of cognitive value to ethical statements was a familiar objection to verificationist theories of meaning, and it seems to be a fair charge against the "Peircian" reading of the Pragmatic Maxim as well. [30]

The Peircian version of the maxim is not, then, without serious problems, and one might ask whether other readings of the maxim might fare better. James claimed that his maxim was essentially Peirce's, but he also suggested that the maxim should be "expressed more broadly" than Peirce did,[31] and most subsequent writers have taken James to have understated his differences with Peirce. That certainly seemed to be the view of Peirce himself, and his evident dissatisfaction with how his original maxim was developed by James can be seen is his suggestion that his own view (and its "poor little

[25] As Peirce and James stated above, "our conception of these effects is the whole of our conception of the object" (Peirce 1878, p. 132), and "[o]ur conception of these effects, then, is for us the whole of our conception of the object" (James 1898, p. 259, James 1904, p. 124, 1907, p. 29).

[26] James 1907, p. 50.

[27] James explicitly assumes that both positions are equally successful at explaining how things stand now (James 1907, p. 51).

[28] James 1907, pp. 50–1.

[29] James 1907, p. 55. James, however, quickly came to have doubts about this purported equivalence, even without these future considerations (James 1909, p. 103).

[30] Of course, it isn't clear whether Peirce himself intended his maxim to apply to our ethical concepts, since he insisted that pragmatism was "a method of ascertaining the meanings, not of all ideas, but only of what I call 'intellectual concepts', that is to say, of those upon the structure of which, arguments concerning objectivity may hinge" (Peirce 1907, p. 421). That said, some (such as Misak 2000, 2013) have argued that ethical claims can be worked into a Peircian framework.

[31] James 1898, p. 259.

maxim")[32] be renamed "pragmaticism" in order to distinguish it from the "pragmatism" that was by then associated with James.[33]

Indeed, James's pragmatic maxim has often been read in ways that are *radically* different from Peirce's.

3. The 'Activist' Reading

One way to do this is to read James's principle as not proposing a conception of *cognitive* significance at all. Rather, the maxim is just read as a way of sorting those philosophical questions that are worth pursuing from those that are not. As Kitcher recently put it:

> James and Dewey share [with the Logical Positivists] the wish to eliminate "insignificant questions" from philosophy—but the apparent communion of goals depends on a bad pun. 'Significance' for them has nothing to do with semantics, or with a verificationist approach to meaning: they are out to focus philosophy on issues that matter to people.[34]

On this reading, the Pragmatic Maxim is just an injunction for us to move, as Dewey famously put it, from the "problems of philosophers" to "the problems of men",[35] and in which we do not "solve" philosophical problems, "we get over them".[36] The maxim thus comes out as something closer to the sort of principle one finds Rorty applying when he claims that for pragmatists like himself "the traditional questions of metaphysics and epistemology can be neglected", not because they are "devoid of meaning" or "rest on false premises" but because "they have no social utility" since "the vocabulary of metaphysics and epistemology is of no practical use."[37] James might differ from Rorty and Dewey about just how much juice could still be squeezed from traditional philosophical problems,[38] but on this reading, all three would agree that pragmatism involved shifting one's focus to the problems which were of "vital" importance.[39]

There are passages in James that can certainly encourage such a reading, as when he claims:

> [M]ost men instinctively . . . do turn their backs on philosophical disputes from which nothing in the line of definite future consequences can be seen to follow. The verbal and empty character of philosophy is surely a reproach with which we are but too familiar. If pragmatism be true, it is a perfectly sound reproach unless the theories under fire can be shown to have alternative practical outcomes, however delicate and distant these may be. The common man and the scientist say they discover no such outcomes, and if the metaphysician can discern none either, the others certainly are in the right of it, as against him. His science

[32] Peirce 1908, p. 448. [33] Peirce 1905, pp. 334–5.

[34] Kitcher 2012, pp. xii–xiii. (See also Kitcher 2012, p. xvii.)

[35] Dewey 1917, p. 95. (See also Dewey 1925, p. 7.) [36] Dewey 1909, p. 40.

[37] Rorty 2007b, pp. 37–8. Of course, Rorty would certainly not refer to this as "a method".

[38] Unlike Rorty and the positivists, James believed that the traditional philosophical questions did have practical import (see Nagl 2004, p. 20).

[39] This contrasts with Peirce who, at least at times, thought that the application of the Pragmatic Method was particularly inappropriate for such "vital" questions (though see Misak 2013, p. 45).

> is then but pompous trifling; and the endowment of a professorship for such a being would be silly.
>
> (James 1907, p. 52)

Further, such a reading of James is almost as old as pragmatism itself, and can be traced back to Lovejoy's argument that James's maxim ultimately gives us neither the "intellectual meaning" nor the "logical validity" of propositions, but rather their "moral worth" and "human significance".[40] According to Lovejoy, James conflated the question of which topics were worth studying with the question which topics were meaningful at all, and mistakenly transformed "a strong conviction concerning the relative importance of propositions into a logical doctrine concerning the import of propositions."[41]

This "activist" reading makes the maxim's focus on the future seem less problematic, since it is more plausible to claim that disputes that have no consequences for future experience are *idle*, than it is to say that they are *meaningless*.[42] On this reading, the maxim isn't intended to produce anything like reductive definitions or complete accounts of meaning. Rather, the maxim is meant to specify that part of meaning that is useful *for philosophical purposes*, since the point is to get the part of meaning relevant to debate. The "verification transcendent" part of meaning may be there, but not a part of meaning that helps us settle the issue. If we are actually going to debate whether Plato stubbed his toe 18 days before his eleventh birthday, then consequences of the sort the maxim emphasizes will be needed. If no such consequences can be found, debate on the issue is futile.

However, the activist reading makes James's choice of the maxim itself, and what he wants to do with it, fairly mysterious. First of all, the activist reading treats the principle as one for sorting those philosophical problems that are worth pursuing from those that are not, but James seemed to consistently describe pragmatism as a method for solving philosophical disputes, not one for deciding whether or not a dispute was worth pursuing. Secondly, if one wanted to simply switch our focus to problems that made a significant difference to our lives, it's hard to imagine that one would choose to do so with a maxim of the sort that James formulates. One needs a lot more to settle whether a question is "idle" than just whether its answer makes some difference to "somebody, somehow, somewhere and somewhen",[43] since anything that has *any* experiential consequences will make such a difference. If the "activist" reading were what James wanted, the principle would have been a terrible way to express it. Even manifestly trivial questions such as "Is there an even or odd number of grains of sand in my nephew's sandbox?" can be tied to some difference in future experience, so the maxim seems ill suited for separating those questions which are "idle" from those which are not.

None of this is to deny that James might have a good deal of sympathy with the Dewey/Rorty/Kitcher position that we should focus on problems that make a difference

[40] Lovejoy 1908b, p. 56. [41] Lovejoy 1908b, p. 59.

[42] Such a reading stresses passages of James such as "if no future detail of experience or conduct is to be deduced from our hypothesis, the debate between materialism and theism becomes quite idle and insignificant. Matter and God in that event mean exactly the same thing—the power, namely, neither more nor less, that could make just this completed world—and the wise man is he who in such a case would turn his back on such a supererogatory discussion" (1907, p. 52).

[43] James 1907, p. 30.

to our lives: it's only to note that this sympathy wasn't what he was expressing with his Pragmatic Maxim.

4. The Subjectivist Reading

This leads to a surprisingly popular explanation of how James's maxim opens things up from the original "Peircian" version. On this reading of James, he proposes that we take the content of a philosophical view to be the practical consequences not (only) of its being true, but (also) of our *believing it*.[44]

This "subjectivist" reading of the maxim is obviously more forgiving than the Peircian one about which questions would be meaningful. It would, for instance, have an easy time allowing ethical statements to be meaningful, since one's ethical beliefs can affect one's behavior in a fairly straightforward way. While the *truth* of the statement "eating meat is always wrong" might not entail any experiences on our part, it seems clear that *believing* it has practical consequences in terms of what we will and won't do because of that belief. On such an account, "Eating meat is always wrong" would be meaningful at least in part because my believing it would lead me to stop eating meat (or at least eat it less).

In much the same way, the subjectivist reading doesn't have the problem of potentially treating claims like "Plato stubbed his toe 18 days before his eleventh birthday" as meaningless. I'm not sure what difference it could make to my future experience if that claim were true, but it seems pretty clear what sort of differences there could be if I *believed* it. In particular, it would make a difference to what I would answer if asked, "Do you think that Plato stubbed his toe 18 days before his eleventh birthday?"

These small advantages, however, are a consequence of one of the major downsides of the subjectivist reading of the maxim—namely, on such a reading *every* claim is going to not only be meaningful, but also have a distinct meaning. Every belief will have *some* consequences that follow from believing it for *someone*, and no two beliefs will have the same consequences for *everyone*. "John bought a female fox" and "John bought a vixen" have distinct meanings because answering "yes" to the question "Did John buy a fox?" is likely to follow from believing the first, while it may be less likely to follow from believing the second. The subjectivist reading thus allows so much that the maxim is completely stripped of its teeth. James claims that many disputes would collapse into insignificance once subjected to this test,[45] but it seems as if no dispute would be "purely verbal" if the subjective effects of believing were allowed to determine the contents of the beliefs involved.

Given how problematic the resulting view is, it is surprising how little textual evidence there is that James endorsed this more "subjective" interpretation of the maxim. Indeed, when James presents the maxim, the formulations he gives invariably favor a reading that

[44] For recent versions of this interpretive strain, see, for instance, Bacon 2012, p. 28, De Waal 2005, p. 21, Hookway 2010, p. 8, Misak 2013, p. 58, Rescher 2000, p. 9, Suckiel 2006, p. 33, Tallise and Aiken 2008, pp. 11–13. (It is also suggested by Brandom's claim that for James, Peirces's principle, amounts to the claim that "the meaning of a claim is the difference that adopting it would make to what one does" (Brandom 2011, p. 20).)

[45] James 1907, p. 30.

ties content to the practical effects of the sentence's *truth* rather than our *believing* it.[46] By contrast, there are no explicit formulations of the maxim where James states, or even suggests, that the consequences of *believing* the sentence to be the relevant ones.

So what reason could there be for ascribing the subjective reading of the maxim to James? Lovejoy first suggested such a reading in 1908, and (at least partially because it fits other preconceptions readers would have about James),[47] many subsequent writers on James seemed happy to just follow Lovejoy's lead on this issue. In that article, Lovejoy argues[48] that James must have something like the subjectivistic version of the maxim in mind, since (1) James takes absolute idealism[49] not to involve any predictions about our future experiences:

> [the Absolute] remains supremely indifferent to what the *particular* facts in our world actually are. Be they what they may, the Absolute will father them. . . . You cannot redescend into the world of particulars by the Absolute's aid, or deduce any *necessary* consequences of *detail* important for your life from your idea of his nature.
>
> (James 1907, p. 40, italics mine)

and (2), James suggests that claims about the Absolute are still meaningful because of the emotional and spiritual comfort they bring:

> the use of the Absolute is proved by the whole course of men's religious history . . . it is indeed not a *scientific* use, for we can make no *particular* deductions from it. It is emotional and spiritual altogether.
>
> (James 1907 p. 131, italics mine)

However, the focus of these passages that Lovejoy appeals to is on a lack of detailed *particular* predictions, not a lack of consequences at all,[50] and the mere fact that there are no particular "scientific" deductions of future experiences that follow from our commitment to the Absolute does not entail that we wouldn't still be committed to there being a concrete difference in *some* (as yet unspecified) future experience. When we look what James actually says about the Absolute right after he claims that we can't deduce any particular consequences from it, he clearly commits himself to the truth of the belief in the Absolute entailing a difference in what would be experienced—it's just that there are no details about either the timeframe or the general form of the good turn of events entailed, and thus no specific predictions. As he puts it:

> What do believers in the Absolute mean by saying that their belief affords them comfort? They mean that since in the Absolute finite evil is "overruled" already, we may, therefore, whenever

[46] See, for instance, James 1902a, p. 351, James 1902b, p. 94, James 1907, pp. 28, 29, 30, James 1909, p. 37, James 1910.

[47] James wrote extensively about the practical consequences of our beliefs, and it is a common understanding of his "The Will to Believe" (James 1896a) to take it to focus on the practical benefits that come from *believing* in God (in contrast to Pascal who focused on the practical benefits associated with the belief's *truth*), and so it can be tempting to read such concerns of his into his understanding of the Pragmatic Maxim.

[48] Lovejoy 1908a, p. 9.

[49] The "absolute" being shorthand for the views James attributes to "absolute idealists" like Royce and Bradley who took the world to be fundamentally a single whole that was, ultimately organized in the best of all possible ways (see Royce 1885, Bradley 1893).

[50] For a good discussion of this, see Sukiel 1982, ch. 3.

> we wish, treat the temporal as if it were potentially the eternal, be sure that we can trust its outcome, and, without sin, dismiss our fear and drop the worry of our finite responsibility. In short, they mean that we have a right ever and anon to take a moral holiday, to let the world wag in its own way, feeling that its issues are in better hands than ours and are none of our business.
>
> (James 1907, p. 41)

Now it seems from this passage James's position is that while we can't *deduce* any *particular* consequences from our belief in the Absolute, we can take "moral holidays" if we believe in it because we can infer that things will work out for the best no matter what we do ourselves. The psychological consequences of believing it are important, but they follow from a conclusion about what the total sum of experiences will be like if the hypothesis is true.

That said, most of the problems with the subjective reading discussed above stem from the fact that it focuses on the effects that a belief *does* have on a believer's behavior, rather than what effects it *should* have.[51] If the subjective maxim were read a more normative way, many of the problems with the descriptive version of the subjective maxim would disappear. For instance, how we *should* behave if we believed that there was a vixen in the forest is no different than how we should behave if we believed that there was a female fox there, even if the actual behavior produced was different.

Such a normative reading would also allow one to avoid another serious problem for the descriptive version of the subjective interpretation of the maxim—namely, that it seems to make it hard to attribute any general content to a philosophical view because just what a belief will lead someone to do varies from person to person. For instance, while James argues that a belief in the Absolute can justify an occasional moral holiday, he couldn't (if the descriptive–subjective reading were right) claim that the content of the Absolute is such a right to take moral holidays unless he had grounds for thinking that this was the actual effect that believing in the Absolute had for most people. If people mistakenly thought that the Absolute required moral seriousness on their part and behaved accordingly, then that would be the content for them. On the other hand, with the more normative reading of the subjective version of the maxim, James can claim that the content of the Absolute is that we can take moral holidays because that is what we *should* be able to do if we believed in the Absolute, even if such a belief didn't actually incline us to take such holidays.

This normative reading of the subjective version of the maxim still runs into the problem that James did seem to focus on the consequences of a claim's *truth* rather than the consequences of our *believing* it, but it does point towards yet another interpretation of the maxim, one that avoids that problem while incorporating much of what the normative version of the subjective interpretation of the maxim hoped to capture.

5. The Practical Reading

While the subjectivist reading has serious problems, there is something to its treating the maxim as tying meaning to our forthcoming actions as well as our sensations, and there are ways of doing this without moving to the subjectivist reading itself.

[51] See Gale 2010, p. 110 for a useful discussion of the "normative" vs. the "causal" phrasings of the maxim in James.

In particular, when James speaks of "what effects of a conceivably practical kind the object may involve", he refers not only to "what sensations we are to expect from it", but also to "what reactions we must prepare."[52] Now, it is a real question what this second aspect is supposed to add to the first. When he summarizes the Principle in *The Varieties of Religious Experience*, this second aspect is cashed out in terms of "what conduct it is fitted to produce" and "what conduct we must prepare in case the object should be true", and while it is easy to conclude that every difference in predicted sensation will produce a difference in the set of actions called for, it is conceivable that there could be a difference in actions called for without there being a difference in predicted sensation. For instance, it is easy to read this second aspect as allowing room for normative claims to have meaning because their truth is "fitted" to produce conduct that is in line with them. "Eating meat is wrong" has a different meaning from "Eating meat is right" because "the conduct we must prepare" with the former precludes eating meat, while the "conduct we must prepare" with the latter includes the possibility of eating it. The second type of practical effect that James mentions in his Pragmatic Maxim thus makes the view more forgiving than both Peirce's version or the positivist criterion of meaning.[53]

The resulting view is "normative" in that it focuses on what the belief is "fitted" to produce. That is, it focuses on what we *should* do if the belief were *true*, not simply on what we might *actually* do if we *accept it* as true.[54] Still, it should also be clear why emphasizing this sort of practical consequence could make one's position sound like the subjective version of the maxim. The actions one should engage in if a sentence were true often are the actions one should engage in if you believed it, and so an appeal to those actions or attitudes can look as if one is appealing to the effects of *believing* the sentence rather than any practical consequence that follows from its *truth*. For instance, in the example of the Absolute that has been given, the view would be that a "practical effect" of the *truth* of the claims associated with the Absolute would be that we are *entitled* to take moral holidays. There is nothing subjectivist about this, but it is easy to confuse with the subjectivist view that understands the meaning of the Absolute as stemming from the fact that we will be more likely to take moral holidays if we *believe* in it. On this reading, if the truth of absolute idealism didn't legitimize moral holidays, then the fact that people who believed in it *actually* felt entitled to such holidays would be irrelevant to its meaning.

[52] James 1898, p. 259, James 1904, p. 124, 1907, p. 29. This is perhaps expressed most clearly in the definition of pragmatism James provides for *Baldwin's Dictionary*: "The doctrine that the whole meaning of a conception expresses itself in practical consequences, consequences either in the shape of conduct to be recommended, or in that of experiences to be expected, if the conception be true" (James 1902b, p. 94).

[53] It would also underwrite James's insistence (James 1909, p. 103) that the pragmatist must recognize the difference between a "spiritually animated maiden" and an "automated sweetheart" (a "soulless body" that was "absolutely indistinguishable" from the former), since the reactions appropriate to one need not be appropriate to the other.

[54] Of course, just what conduct is recommended by the truth of a claim will (as stressed in Brandom 2011, p. 50) depend on what the individual desires, and this will produce, not only a more expansive holism than the sort associated with the Peircian version, but will also be more individualistic, since the desires in question will vary from individual to individual more than the predicted experiences would.

We can see how this works in James's discussion of the Eucharist. While Peirce explicitly states that the Pragmatic Maxim shows that debates about the Eucharist are without significance,[55] James seems to take the opposite line, namely:

> Substance here would appear to have momentous pragmatic value. Since the accidents of the wafer don't change in the Lord's supper, and yet it has become the very body of Christ, it must be that the change is in the substance solely. The bread-substance must have been withdrawn, and the divine substance substituted miraculously without altering the immediate sensible properties. But tho these don't alter, a tremendous difference has been made, no less a one than this, that we who take the sacrament, now feed upon the very substance of divinity. The substance-notion breaks into life, then, with tremendous effect.
>
> (James 1907, pp. 46–7)

James notably, and rather unhelpfully, doesn't say what the "tremendous effect" in this case is supposed to be. Since the accidents are the same, it is often assumed that the "practical effects" of the switch in substance cannot be objective, and this has led to a reading of Peirce and James in which Peirce takes talk of transubstantiation to be meaningless because no experiential consequences follow from its *truth,* while James takes such talk to be meaningful because of the mental comfort and satisfaction that comes from *believing* in it. However, on the reading suggested here the difference comes, not from the psychological effects of believing in transubstantiation, but rather from the acts and attitudes normatively required by the bread and wine becoming "the very substance of the divinity". Differing attitudes towards the Host's desecration, whether one should genuflect to the Host on the altar upon entering the church, how one should treat extra communion wafers can seem to be demanded by the *truth* of the various views of transubstantiation, and this accounts for their difference in meaning for James.

The practical reading of the maxim thus allows for pairs of claims to have distinct meanings, even if they entail the same future experiences,[56] but the added fineness of grain that comes from focusing on "the conduct we must prepare" is also a vital part of determining the meaning of our philosophical claims for James, because he believes that, just as much as with the sensations predicted, the more "practical" side of a philosophical position will ultimately determine whether it is acceptable or not. As he puts it early on in his "The Sentiment of Rationality", if two conceptions of the world are equally consistent and both account for the available evidence, "that one which awakens the active impulses, or satisfies other æsthetic demands better than the other, will be accounted the more rational conception, and will deservedly prevail".[57]

Indeed, James often ties his discussion of various philosophical questions that may seem at bottom "metaphysical" (free will vs. determinism, materialism vs. theism, monism vs. pluralism) to the practical conclusions that one should draw from such competing positions (e.g. giving up hope if materialism is true, being complacent if idealism is true). Such an argumentative gambit can often seem like engaging in a kind of wishful thinking

[55] Peirce 1878, pp. 131–2.

[56] However, James's independent commitment to radical empiricism (particularly its postulate that "the only things that shall be debatable among philosophers shall be things definable in terms drawn from experience" (James 1909, p. 6)) may temper his ability to appeal to this aspect of the maxim.

[57] James 1882, p. 66. One should note that, in spite of his criticism of James's views on the "The Will to Believe", Peirce arguably moves in this direction when he claims that the "ultimate test" of a

(drawing metaphysical conclusions from what one wants to be true), but it may just as well be that James is arguing that in these cases (which he consistently claims to be empirically underdetermined), such practical considerations, whether we like it or not, often determine the views we adopt. Focusing merely on the experiences predicted, rather than the more normative expectations, however, obscures these deciding factors and makes the debates seem more intractable than they really are. Spelling out philosophical views in terms of the practical version of the Pragmatic Maxim puts these practical considerations, which are always (even if just subconsciously) motivating, into the foreground so that they can be targets of rational scrutiny as well.

This aspect of James's method, that of making the subjective engines driving our philosophical views explicit (and thus subject to rational evaluation) is far removed from simply endorsing wishful thinking. Rather, it involves recognizing that such "subjective" factors will inevitably affect what views we ultimately adopt, and that it is thus best to have them subject to criticism, since the practical upshot we *actually* draw from a philosophical view may not be the one that we *should* draw from it. (As when the absolute idealist thinks that his or her view underwrites a type of moral seriousness, while James argues that it actually legitimates a type of moral complacency.) If our practical demands help determine which views (among those that pass basic logical muster) we will ultimately accept, making the practical consequences of various philosophical views explicit can go a long way towards settling philosophical disputes, because we can, for instance, (1) see what ultimately makes a view like absolute idealism appealing, and (2) argue for a replacement for it that captures those appealing aspects as well (which James claims his form of meleoristic pluralism does).

One problem[58] with this version of the maxim is that explaining the meaning of our claims in terms of such normative consequences means that the claims themselves can't be used to *explain why* we are committed to doing what we do. It might seem that we should, for instance, genuflect in front of the Host *because* it is literally the body of Christ, but if what distinguishes the meaning of that claim from the more symbolic interpretation of the Eucharist just *is* the set of normative consequences which includes genuflection, not throwing out unused communion wafers, etc., then the attempt to explain *why* we should genuflect in front of the Host in terms of that claim seems circular. (Of course, something like this worry is present for the "Peircian" reading as well, and one might think that the fact that explaining the meaning of "a large meteor landed 10,000 years ago where Toronto is now located" in terms of future experiences, would similarly preclude our explaining why we would find a crater formation under the soil around Toronto in terms of a meteor having landed there 10,000 years ago.)

One might be able to assuage this worry by claiming that the maxim isn't, strictly speaking, intended to provide meaning equivalences (which was, as a self-standing problem, the sort of issue that James was comparatively unconcerned with), but is just meant to capture the fact that to clearly understand a claim, you need to know what follows from its truth. Indeed, James's and Peirce's shared suggestion that "the whole of our conception of

hypothesis "must lie in its value in the self-controlled growth of a man's conduct in life" (Peirce 1908, p. 446).

[58] Which I'd like to thank an anonymous referee for this volume for stressing.

the object" is "our conception of these [practical] effects"[59] could be understood as compatible with this more modest reading. Even if the maxim didn't produce full meaning-equivalencies, capturing how things would be different if one claim rather than another were true could still do much the same work in helping settle philosophical disputes.

Even so, as a philosophical method, the practical version of the maxim brings with it problems of its own. In particular, the normative matters that determine "the conduct we must prepare" can be controversial, often more so than the views that they are supposed to explain, and so using such predictions to elucidate the meaning of a controversial subject can seem quixotic at best. Indeed, by the time he published *The Meaning of Truth*, it was quite clear to James that most, if not all, of his interlocutors did not accept his proposed explications of the meanings of their views.[60] What James took to be the essential practical upshot of various philosophical views were taken to be at best peripheral by their defenders.

After all, the normative consequences that come with, say, the different cosmologies James considers are themselves far from obvious. Not only because the empirical consequences of such cosmologies are hard to predict, but also because how one should react even if one could be sure of those consequences is open to debate. For instance, when James claims that "Materialism means simply the denial that the moral order is eternal, and the cutting off of ultimate hopes",[61] he is working with particular assumptions about what a materialist cosmology must look like (namely, that all of the stars must burn out and all life die out), and someone who didn't share such assumptions wouldn't view "materialism" as having such a meaning at all. The pessimism comes not, then, from materialism itself, but from certain (admittedly plausible, but ultimately optional) assumptions about what a scientific cosmology must predict for our future.

Furthermore, even if he could be assured that materialism did entail that all conscious life would permanently die out, James's pessimistic and somewhat depressive attitude towards the universe burning out in the far, far distant future isn't shared by all,[62] and deciding the question of whether there is any point to our actions if no one will be there to remember them in a million years isn't obviously an easier question to settle than the question of whether God created our world. While there is much to be said for making the practical consequences of our views as explicit as possible, just what those practical consequences are is often itself subject to philosophical dispute.

In conclusion, then, each version of the Pragmatic Maxim brings with it a range of problems, so it should not be surprising that while there are many philosophers today who identify themselves as "pragmatists", it is not because they endorse any version of the method that Peirce and James initially identified with the view.

[59] Peirce 1878, p. 132, James 1898, p. 259, James 1904, p. 124, 1907, p. 29.

[60] James 1909, p. 5.

[61] James, 1907, p. 55.

[62] It is more characteristic of what he calls the "sick soul" rather than the "healthy minded" (see James 1902a, ch. 4–7), and his assertion that the sick soul simply has a deeper and wider appreciation of reality (James 1902a, 136–8) is largely undefended. He mentions that only the sick soul is aware of the evils that are really there, but gives no reason to think that the healthy minded couldn't be aware of such evils as well (other than that they are not as bothered by them as the sick souls are).

References

Ayer, A. J. 1968. *The Origins of Pragmatism*. San Francisco: Freeman, Cooper & Company.

Bacon, M. 2012. *Pragmatism: An Introduction*. Cambridge: Polity Press

Bernstein, R. 1999. "American Pragmatism: The Conflict of Narratives", in Saatkamp, (ed.) *Rorty and Pragmatism*. Nashville: Vanderbilt University Press. pp. 54–67.

Bernstein, R. 2010. *The Pragmatic Turn*. Cambridge: Polity Press.

Bradley, F. H. 1893 *Appearance and Reality*. New York: Macmillan.

Brandom, R. 1994. *Making it Explicit*. Cambridge: Harvard University Press.

Brandom, R. 2000. *Articulating Reasons*. Cambridge: Harvard University Press.

Brandom, R. 2011. *Perspectives on Pragmatism*. Cambridge: Harvard University Press.

Carnap, R. 1936. "Truth and Confirmation". Translated and reprinted in Feigl and Sellars, 1949. pp. 119–27.

Carus, P. 1911. *Truth on Trial: An Exposition of the Nature of Truth*. Chicago: The Open Court Publishing Company.

De Waal, C. 2005. *On Pragmatism*. Belmont: Wadsworth.

Dewey, J. 1909. "The Influence of Darwinism on Philosophy". Reprinted in McDermott (ed) *The Philosophy of John Dewey*. Chicago: University of Chicago Press, 1973, 1981. pp. 31–40.

Dewey, J. 1917. "The Need for a Recovery of Philosophy". Reprinted in McDermott (ed) *The Philosophy of John Dewey*. Chicago: University of Chicago Press, 1973, 1981. pp. 58–97.

Dewey, J. 1925. *Experience and Nature*. Chicago: Open Court.

Gale, R. 2010. "The Deconstruction of Traditional Philosophy in William James's Pragmatism", in J. Stuhr (ed.) *100 Years of Pragmatism: William James's Revolutionary Philosophy*. Bloomington and Indianapolis: Indiana University Press, 2010. pp. 108–23.

Feigl, H. and Sellars, W. 1949. *Readings in Philosophical Analysis*. New York: Appleton-Century-Crofts.

Haack, S. 2006. "Not Cynicism, but Synechism: Lessons from Classical Pragmatism", in Shook and Margolis, (eds.) *A Companion to Pragmatism*. Oxford: Blackwell. pp. 141–53.

Hookway, C. 2004. "The Principle of Pragmatism: Peirce's Formulations and Examples". *Midwest Studies in Philosophy*, vol. 28, 2004: 119–36. Reprinted in his *The Pragmatic Maxim*, New York, Oxford University Press, 2012.

Hookway, C. 2010. "Pragmatism", in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2010 Edition), URL = <http://plato.stanford.edu/archives/spr2010/entries/pragmatism/> (accessed September 17, 2015).

James, W. 1882. "The Sentiment of Rationality". Reprinted in James 1896b, 1979.

James, W. 1896a. "The Will to Believe". Reprinted in James 1896b, 1979.

James, W. 1896b, 1979. *The Will to Believe and Other Essays in Popular Philosophy*. Cambridge: Harvard University Press.

James, W. 1898, "Philosophical Conceptions and Practical Results". Reprinted in James 1907, 1979.

James, W. 1902a, 1985. *The Varieties of Religious Experience*. Cambridge: Harvard University Press.

James, W. 1902b. "Pragmatism". Entry in *Baldwin's Dictionary*. Reprinted in his *Essays in Philosophy*, Cambridge: Harvard University Press. 1978. p. 94.

James, W. 1904. "The Pragmatic Method". Reprinted in James 1978. pp. 123–39.

James, W. 1907, 1979. *Pragmatism*. Cambridge: Harvard University Press.

James, W. 1909, 1975. *The Meaning of Truth*. Cambridge: Harvard University Press.

James, W. 1910, 1979. *Some Problems of Philosophy*. Cambridge: Harvard University Press.
James, W. 1978. *Essays in Philosophy*. Cambridge: Harvard University Press.
Kitcher, P. 2012. *Preludes to Pragmatism: Towards a Reconstruction of Philosophy*. New York: Oxford University Press.
Lewis, C. I. 1934. "Experience and Meaning". *The Philosophical Review,* 43, 1934, pp. 125–46. Reprinted in Feigl and Sellars, 1949, pp. 128–45.
Lovejoy, A. O. 1908a. "The Thirteen Pragmatisms". Reprinted in his *The Thirteen Pragmatisms and Other Essays*. Baltimore: Johns Hopkins University Press, 1963, pp. 1–29.
Lovejoy, A. O. 1908b. "Pragmatism and Theology". Reprinted in his *The Thirteen Pragmatisms and Other Essays*. Baltimore: Johns Hopkins University Press, 1963, pp. 40–78.
Malanchowski, A. 2010. *The New Pragmatism*. Montreal & Kingston: McGill-Queens University Press.
Menand, L. 2001. *The Metaphysical Club.* New York: Farrar, Straus and Giroux.
Misak, C. 1995. *Verificationism: Its History and Prospects*. New York: Routledge.
Misak, C. 2000. *Truth, Politics, Morality: Pragmatism and Deliberation*. New York: Routledge.
Misak, C. 2004. "Making Disagreement Matter: Pragmatism and Deliberative Democracy." *Journal of Speculative Philosophy* 18.1: 9–22. Reprinted in Talisse and Aiken (eds.) *The Pragmatism Reader*. Princeton: Princeton University Press, 2011.
Misak, C. 2007. "Introduction" to C. Misak (ed.) *New Pragmatists*. New York: Oxford University Press, 2007, pp. 1–6.
Misak, C. 2010. "Richard Rorty's Place in the Pragmatist Pantheon", in Randall E. Auxier & Lewis Edwin Hahn (eds.) *The Philosophy of Richard Rorty* (vol 32 of the *Library of Living Philosophers)*. Chicago: Open Court, 2010, pp. 27–43.
Misak, C. 2013. *The American Pragmatists*. New York: Oxford University Press.
Mounce, H. O. 1997. *The Two Pragmatisms*. London: Routledge.
Nagl, L. 2004. "The Insistence on Futurity: Pragmatism's Temporal Structure", in Egginton and Sandbothe (eds.) *The Pragmatic Turn in Philosophy*. Albany: State University of New York Press, pp. 11–29.
Peirce, C. S. 1868a. "Questions Concerning Certain Faculties Claimed for Man". Reprinted in Peirce 1992, pp. 11–27.
Peirce, C. S. 1868b. "Some Consequences of Four Incapacities". Reprinted in Peirce 1992, pp. 28–55.
Peirce, C. S. 1878. "How to Make Our Ideas Clear". Reprinted in Peirce 1992, pp. 124–41.
Peirce, C. S. 1905. "What Pragmatism Is". Reprinted in Peirce 1998, pp. 331–45.
Peirce, C. S. 1907. "Pragmatism", Reprinted in Peirce 1998, pp. 398–433.
Peirce, C. S. 1908. "The Neglected Argument for the Reality of God", Reprinted in Peirce 1998, pp. 434–50.
Peirce, C.S. 1975–1987. *Charles Sanders Peirce: Contributions to The Nation* (four volumes). K. L. Ketner & J. E. Cook, eds, Lubbock: Texas Tech University Press.
Peirce. C. S. 1992. *The Essential Peirce, Volume 1.* Bloomington and Indianapolis: Indiana University Press.
Peirce. C. S. 1998. *The Essential Peirce, Volume 2.* Bloomington and Indianapolis: Indiana University Press.
Price, H. 2011. *Naturalism without Mirrors*. New York: Oxford University Press.
Price, H. 2013. *Expressivism, Pragmatism and Representationalism*. New York: Cambridge University Press.
Putnam, H. 1981. *Reason, Truth and History*. New York: Cambridge University Press.

Putnam, H. 1994. "Pragmatism and Moral Objectivity". In his *Words and Life*. Cambridge: Harvard University Press, 1994, pp. 151–81.
Putnam, H. 1995. *Pragmatism: An Open Question*. Cambridge: Blackwell.
Putnam, H. 2002. *The Collapse of the Fact/Value Dichotomy*. Cambridge: Harvard University Press.
Quine, W. V. 1951. 'Two Dogmas of Empiricism". Reprinted in Quine, W. V. *From a Logical Point of View*. Cambridge: Harvard University Press, 1953, pp. 20–46.
Reichenbach, H. 1938. *Experience and Prediction: An Analysis of the Foundation and the Structure of Knowledge*. Chicago: University of Chicago Press.
Rescher, N. 2000. *Realistic Pragmatism*. Albany: State University of New York Press.
Rorty, R. 1979. *Philosophy and the Mirror of Nature*. Princeton: Princeton University Press.
Rorty, R. 2004. "A Pragmatist View of Contemporary Analytic Philosophy". In Rorty 2007a, pp. 133–46.
Rorty, R. 2007a. *Philosophy as Cultural Politics*. New York: Cambridge University Press.
Rorty, R. 2007b. "Main Statement by Richard Rorty", in R. Rorty & P. Engel (eds) *What's the Use of Truth?* New York: Columbia University Press, 2007, pp. 31–46.
Royce, J. 1885. *The Religious Aspect of Philosophy*. New York: Houghton, Mifflin and Company.
Russell, B. 1908. "Transatlantic Truth". *The Albany Review*, January 1908. Reprinted as "William James' Conception of Truth" in his *Philosophical Essays*. London: Routledge: 1910/1994, pp. 112–30.
Soames, S. 2016. "Methodology in Nineteenth and Early Twentieth-Century Analytic Philosophy", in H. Cappelen, T. Gendler, and J. Hawthorne, *The Oxford Handbook of Philosophical Methodology*, Oxford: Oxford University Press, chapter 3.
Suckiel, E. K. 1982. *The Pragmatic Philosophy of William James*. Notre Dame: University of Notre Dame Press.
Suckiel, E. K. 2006. "William James", in J. Shook & J. Margolis (eds) *A Companion to Pragmatism*. Oxford: Blackwell, pp. 30–43.
Talisse, R. B. and Aikin, S. F. 2008. *Pragmatism: A Guide for the Perplexed*. New York: Continuum.

PART III

TOPICS

CHAPTER 12

REFLECTIVE EQUILIBRIUM

YURI CATH

1. Introduction

The method of reflective equilibrium is a method for figuring out what to believe about some target domain of philosophical interest like, say, justice, morality, or knowledge. This method is most closely associated with John Rawls who introduced the term 'reflective equilibrium' into the philosophical lexicon in *A Theory of Justice* (1971), where he appealed to this method in arguing for his famous theory of justice as fairness. However, Rawls had already advocated essentially the same method (without calling it 'reflective equilibrium') in an earlier paper (Rawls 1951), and Goodman (1954) is widely identified (including by Rawls himself) as being the first clear advocate of this kind of method, albeit with respect to a different normative domain, namely, logic.

Due to Rawls' influence, there has been a great deal of discussion of reflective equilibrium (or 'RE') in moral philosophy, and many philosophers have suggested that this method is uniquely suited to theorizing about morality or normativity more generally. However, the method of RE has also been claimed to be a method that plays a central role in all areas of philosophical inquiry, including those that are concerned with non-normative subjects, like metaphysics. This idea that the method of RE is employed in all areas of philosophy has been endorsed not only by proponents of this method (see e.g. Lewis 1983, pp. x–xi), but also by critics who agree that this method is used in this way, but who argue that this is a practice that needs to be abandoned or reformed (see e.g. Cummins 1998; Stich 1990; Weinberg, Nichols, and Stich 2001).

As well as different views of the scope and status of this method, there are also different views of what the method is, to the point where it can be misleading to talk (as is standard) of *the* method of RE. Accordingly, my first aim in this chapter will be to expose some of the subtleties involved in interpreting this method (§§2–4.3). I will then go on to consider some of the main objections to RE (§5). In closing, I will make some remarks about how the method of RE relates to recent debates about the role of intuitions in philosophy (§6).

2. An Initial Sketch

The aim of this section is to offer an initial sketch of the method of RE which will serve as a useful basis for our discussion. Following Scanlon (2003) and others, the method of RE can be usefully conceptualized as involving three distinct stages:

Stage 1: In this stage one identifies a relevant set of one's initial beliefs (or judgements or intuitions) about the relevant domain. These initial beliefs are often characterized as concerning particular rather than general features of the relevant domain. For example, if the domain is justice then one's set of initial beliefs might include the belief that a particular action was just, or if the domain is logic it might include the belief that a particular token inference was invalid, or if the domain is knowledge it might include the belief that a particular subject possesses knowledge, and so on. However, proponents of the method of RE often allow that the set of initial beliefs can also include beliefs about more general features of the relevant domain, including abstract general principles (see e.g. Rawls 1974, p. 8).

Stage 2: In the second stage one tries to come up with an initial set of theoretical principles that would systematize or account for the initial beliefs identified at the first stage. Scanlon (2003, pp. 140–1), in describing Rawls' version of the method of RE, writes that one is trying to come up with principles such that "had one simply been trying to apply them rather than trying to decide what seemed to be the case as far as justice is concerned, one would have been led to this same set of [initial] judgments".[1]

Stage 3: There are likely to be conflicts between one's initial beliefs and one's initial principles that aim to account for those beliefs. Furthermore, there are also likely to be conflicts between the members of one's initial set of beliefs, and perhaps even between the members of one's initial set of theoretical principles. These conflicts lead to the need for a third stage in which one engages in a reflective process of moving back and forth between these two sets and eliminating, adding to, or revising the members of either set until one ends up with a final set of beliefs and principles which cohere with each other. This final state is called a state of reflective equilibrium.

One point worth clarifying is that what the theoretical principles are meant to account for, and be in equilibrium with, is not the psychological fact that one has certain initial beliefs. This point is often obscured because: (i) descriptions of this method typically just speak of 'accounting for our beliefs', or 'bringing our beliefs into equilibrium'; and (ii) such descriptions are ambiguous, given that propositional attitude terms like 'belief' can pick out both the psychological state of believ*ing* and the proposition that is believ*ed* to be true (as is the case for 'intuition' and 'judgement'). However, on close inspection, it is almost always clear that such descriptions should be interpreted in a non-psychologistic way. That is, when proponents of RE talk of theories 'accounting for our initial beliefs', they should

[1] Sometime the method of RE is described in ways that do not fit so well with the characterization I have given of this second stage. For example, I said that in the second stage one tries to 'come up with' theoretical principles that would account for one's initial beliefs, which suggests that one need not have believed these principles to be true prior to engaging in this reflective process. But sometimes the method is described so that the theoretical principles one identifies are just further things that one already believed to be the case before applying this method.

be interpreted as making a claim equivalent to something like 'accounting for the assumed truth of the contents of our initial beliefs', as opposed to 'accounting for the fact that we have certain initial beliefs'.[2] Similarly, when proponents of RE talk of 'our beliefs being in a state of equilibrium', such claims are best interpreted as being equivalent to something like 'the contents of our beliefs being in a state of equilibrium' or 'our beliefs being in a state of equilibrium in virtue of their contents being in a state of equilibrium'. Some proponents of the method of RE do explicitly clarify this interpretative point,[3] but often it is left merely implicit, which can lead to unnecessary confusion and to misplaced criticisms.[4]

With that clarification in mind, we can explain how this method is meant to be one of figuring out what to believe about the target domain. The idea is that the process of bringing the contents of one's beliefs and one's theoretical principles into a state of equilibrium is one that should be mirrored by corresponding changes in one's belief states, so that by the end of this process the contents of one's resulting beliefs about the relevant domain should be captured by the final coherent set of propositions that one reaches in stage 3.

Importantly, proponents of this method usually add the qualification that this state of equilibrium is an ideal that we should strive towards but will perhaps never achieve. For this reason, RE is best viewed as a method that one is meant to continuously return to and reapply, rather than as a method that one would apply once and then set aside.

3. Interpreting the Sketch

To help fill in our initial sketch it will be useful to now consider a series of questions about how to interpret it.

What is meant to recommend this method of forming beliefs about the target domain? Proponents of this method hold that, when applied correctly, it will lead to beliefs that enjoy some positive normative status, where the most common suggestion is that these beliefs will be *justified* (or, at least, that one will have a justification to so believe). And proponents of this method often go further and suggest that it is the *best* and perhaps even the *only* method by which we can form justified beliefs about the relevant domain. For example,

[2] To see this point, it is useful to consider examples of the kinds of conflicts between our initial beliefs or intuitions and our initial theories that are meant to be resolved by the method of RE. Consider a familiar case from epistemology, namely, the conflict between the justified true belief ('JTB') analysis of knowledge and our intuition that the Gettier subject has non-knowledgeable justified true belief ('NKJTB'). This is a paradigm example of the kind of conflict between 'intuition' and 'theory' that the method of RE is meant to address. But, as Williamson (2007, p. 245–6) points out, it makes no sense if we interpret this conflict as one between the JTB analysis and the psychological fact that we believe or intuit that the Gettier subject has NKJTB because the assumed truth of the JTB analysis is consistent with that psychological fact.

[3] For example, Sayre-McCord (1996) makes this point clear when he writes: "The relative coherence of a set of beliefs is a matter of whether, and to what degree, the set exhibits (what I will call) *evidential consistency, connectedness,* and *comprehensiveness*…. Each … is a property of a set of beliefs, if it is at all, only in virtue of the evidential relations that hold among the **contents** of the beliefs in the set" (p. 166 bold emphasis added).

[4] For example, see the arbitrariness objection discussed in §5.2.

Scanlon (2003, p. 149) endorses both these claims with respect to morality and other (non-specified) subjects:

> [I]t seems to me that this method, properly understood, is in fact the best way of making up one's mind about moral matters and about many other subjects. Indeed, it is the only defensible method: apparent alternatives to it are illusory.

Why think that the beliefs formed by this method would be justified? The method of RE is standardly interpreted as relying on a coherentist theory of justification and, indeed, is often referred to as simply being a 'coherence method'. This coherentist interpretation of the method of RE is sometimes disputed for reasons that we will discuss in §5.3. But for now it will suffice to point out how this standard interpretation, if correct, provides a straightforward answer to the justification question, namely, that any belief formed by this method will be justified simply in virtue of it being a member of a system of beliefs that cohere with each other. On this interpretation then, the method of RE is minimally committed to the claim that a belief's being part of a coherent system of beliefs is a *sufficient* condition for it being a justified belief.[5] And when proponents of RE suggest that this method is the *only* way of reaching justified beliefs about some relevant domain, they appear to commit themselves to the idea that it is a *necessary* condition of a belief's being justified (at least for beliefs about the relevant domain) that it be a part of a system of beliefs that is coherent to some degree.[6]

What exactly does it mean to say that the beliefs one reaches at step 3 'cohere' with each other or are in a state of 'equilibrium'? Proponents of RE often do not provide much in the way of detailed answers to this question. One can find more detailed answers in the coherentism literature but there is no consensus account of what coherence is, and it is widely acknowledged that existing accounts are inadequate in different ways.[7] However, for our purposes, it will suffice to note that common to almost all accounts of coherence (in both the coherentist and RE literatures) is the very general thought that increasing the coherence of one's belief system is (at least partly) a matter of minimizing conflicts between the contents of one's beliefs, and maximizing certain relations of support between those contents. These notions of conflict and support are then analysed in a variety of ways by appealing to some mixture of deductive, probabilistic, evidential, or explanatory, relations between the contents of one's beliefs.[8]

[5] Or, alternatively, instead of appealing to this idea that it is sufficient for one's belief being doxastically justified, one might merely appeal to the weaker claim that it is a sufficient condition for having propositional justification to so believe.

[6] I say 'to some degree' in relation to the fact that, as noted earlier, the state of reflective equilibrium is usually thought of as an ideal that we should aim for but may never reach. But proponents of the method of RE will obviously want to say that we can still justify our beliefs (at least to some degree) by making steps towards this ideal.

[7] For example, Bonjour (1985) offered one of the most prominent and detailed accounts of the nature of coherence, but even he saw his account as "a long way from being as definitive as desirable" (1985, p. 101).

[8] How exactly should one bring one's beliefs into a state where they cohere with each other? For example, suppose that one's initial set of moral beliefs includes the belief that it would be morally impermissible for a surgeon to save the lives of five of their patients by giving them the organs of one of their other patients against the wishes of that patient (Foot 1967, Thomson 1976). Furthermore, suppose that one's initial set of moral principles include some simple act-consequentialist principle

What constraints, if any, are placed on the initial set of beliefs one identifies at stage 1? Some prominent proponents of this method—like Goodman (1953) and Lewis (1983)—place very few, if any, constraints on this set of initial beliefs. On the other hand, Rawls placed very specific constraints on the judgements that are the initial inputs into his version of the method of RE. According to Rawls, we should begin with only our "considered judgments", where he uses this as a technical term for those judgements which satisfy a range of constraints aimed at eliminating "judgments [that] are likely to be erroneous or to be influenced by excessive attention to our own interests" (Rawls 1971, p. 42). Rawls' version of RE can be thought of as one on which there is an intermediate step between stages 1 and 2 where one checks one's initial set of gathered judgements and filters out any that do not meet his constraints. These include the constraints that these judgements should not be ones made when one is upset or frightened, or where one's self-interests could be impacted by what the answer is to the relevant question. Rawls also requires that we only include those judgements in which we are confident, and which will be held stably over time. Another restriction that is sometimes placed on the initial inputs identified at stage 1 is that they have to be *intuitive* judgements or beliefs, and sometimes these inputs are described as simply being intuitions. Rawls (1951, p. 183) endorses a restriction of this kind, but it is worth noting that that he has a very minimal sense of this restriction in which an intuitive judgement is simply one that is not "guided by a conscious application of principles so far as this may be evidenced by introspection".

4. Many Methods

We have already indicated how the method of RE might be developed in subtly different ways depending, for example, on what restrictions one places on the initial beliefs identified at stage 1, or how one conceives of the nature of coherence. The aim in this section is to identify three more significant divisions between different interpretations of this method.

4.1 *Deliberative versus Descriptive*

We have been assuming that the method of RE is a method for figuring out *what to believe* about some target domain. Following Scanlon (2003), we can call this *the deliberative interpretation* of this method. My interest in this chapter is just in the deliberative interpretation. But it is worth noting, as Scanlon discusses, that Rawls himself seems to move back-and-forth between this deliberative conception and what Scanlon labels *the descriptive interpretation* of this method. On the descriptive interpretation, the aim of the method

that, if correct, would classify this action as being morally obligatory. How should one resolve this conflict? Should one reject the initial belief or the theoretical principle or both? Again, proponents of the method of RE do not say as much about this kind of issue as one might like. But one idea that is present in many statements of the method of RE is that decisions about how to resolve such questions should be sensitive to the *strength* of one's initial beliefs, as well as the *power* of the theoretical principles. See DePaul (1998, p. 295) for a nice discussion of these ideas.

of RE is to reveal one's implicit *conception* of the relevant domain. For example, RE aims at revealing our conception of, say, morality, as opposed to morality itself.

There may seem to be a tension between these two interpretations of the method of RE. However, Scanlon argues that, for Rawls, the descriptive version of the method depends on the deliberative version of it. This is because the way to reveal one's implicit conception of justice is to first figure out what to believe about justice itself by way of using the deliberative version of RE as described in §2. Whether employing the deliberative version of RE would actually be a good method for revealing one's conception of justice is, to my mind, far from clear. But, for our purposes, it will suffice to have just distinguished the deliberative interpretation from the descriptive, if only to put the latter aside.

4.2 *Narrow versus Wide*

Our initial sketch of the method of RE was a description of what is called the method of *narrow reflective equilibrium* (NRE) as opposed to what is called the method of *wide reflective equilibrium* (WRE). This distinction was first explicitly labelled as such in Rawls (1974), although it is usually thought of as being implicitly present in *A Theory of Justice* (as Rawls himself suggests in his 2001, p. 31), and most proponents of RE endorse the wide version of this method.

The method of RE described in 2.1 is 'narrow' in two relevant senses: (i) it only aims at bringing two sets of things into a state of equilibrium, namely, the set of one's initial beliefs and the set of theoretical principles which are meant to account for those beliefs; and (ii) one's initial beliefs and theoretical principles are both about the same target domain. In contrast to (i), on the method of WRE one is trying to reach a state of equilibrium which will hold between these two sets and a third set of things, namely, any of one's other beliefs or "background theories" (Daniels 1979, p. 258) which are thought to be of some relevance to assessing one's initial beliefs and principles.

These further beliefs may also be about the target domain. For example, when Rawls (1974) introduces the notion of WRE, what he appears to have in mind is an equilibrium between, not only one's considered moral judgements and one's initial moral theories which are meant to account for those judgements, but also one's beliefs about the range of alternative moral theories and the arguments which are meant to support those theories. However, in contrast to (ii), these further beliefs may also be about other domains altogether. For example, in discussions of WRE in moral philosophy it is often suggested that our beliefs about psychology, the theory of meaning, and metaphysics, might all potentially be relevant to figuring out what to believe about morality and, hence, that our idealized aim in moral theorizing should be a state of WRE that includes any such further beliefs insofar as they are relevant.

4.3 *Coherentist versus Foundationalist Interpretations*

As mentioned in §3, the method of RE is standardly viewed as being a coherentist method. However, a number of philosophers have suggested that this method is actually best

interpreted as being committed to some kind of foundationalism. To help frame this issue it will be useful to make a distinction between *strong* versus *moderate* foundationalism, following Bonjour (1985, pp. 26–30). Strong foundationalism is a view that endorses something like the following two claims: (i) there are basic beliefs and (ii) these basic beliefs have some further special epistemic properties that make them a secure foundation for one's non-basic beliefs; where a basic belief is (roughly) a belief that is non-inferentially justified and which can justify other beliefs, and the kind of further epistemic properties that have been historically ascribed to basic beliefs include those of being "*infallible, certain, indubitable*, or *incorrigible*" (Bonjour 1985, pp. 26–30). Moderate foundationalism, on the other hand, is a view on which there are basic beliefs but they do not possess these further epistemic properties.

Proponents of RE do appear to be committed to denying strong foundationalism, that is, at least with respect to those domains to which that they think this method is applicable. For, as we have seen, it is an important feature of the method of RE that *any* of one's beliefs about the relevant domain can, in principle, be rightly overturned on the basis of applying this method. And it is hard to see how one could square that assumption with the idea that some of those beliefs are not possibly false, or that they can't be rationally doubted, etc.

However, one could obviously reject strong foundationalism in this way and still endorse moderate foundationalism, which suggests that the method of RE is at least consistent with foundationalism. Furthermore, a number of commentators—including Ebertz (1993), Holmgren (1989), McMahan (2000), and Pust (2000)—have all suggested that the method of RE is best interpreted as being committed to the idea that the initial beliefs which are the inputs into this method must already be justified to some degree.

Often such claims seem to be motivated by a version of a standard worry about coherentism, namely, that increasing the coherence of one's beliefs can only "amplify" any justification already possessed by one's beliefs, but by itself cannot confer justification on one's beliefs. And, in order to avoid worry, it is sometimes suggested that the method of RE is best interpreted as being committed to some form of epistemic or phenomenal conservatism such that merely believing, or intuiting, that p is, in the absence of defeaters, a sufficient condition for having some degree of justification for believing that *p* (see e.g. Pust 2000, ch. 1).

Interestingly, similar issues arise in the coherentism literature where many self-proclaimed coherentists endorse some form of epistemic conservatism. For example, Lycan (1998) offers a well-known explanatory form of coherentism, and on his view conservatism is one of the main explanatory virtues which can increase the coherence of a belief system. Furthermore, as Poston (2012, p. 78) points out, many historically important defenders of coherentism have endorsed related positions: "Lycan's coherentism falls in line with the explanatory coherentist accounts of Goodman, Quine, Sellars, and Rawls by including a commitment to conservatism, the thesis that the mere holding of a belief confers some epistemic justification on its content."

One might naturally think that what this shows us is that many supposed coherentists are not really coherentists at all, being instead proponents of moderate foundationalism. But this would be a mistake. For example, on Lycan's view, while "a belief is justified by the bare fact of its seeming to be true" (2012, p. 9) it is only so justified to a tiny degree and, crucially, it cannot justify other beliefs without the support of "other beliefs of varying grades

of theoreticity, indeed relative to the subject's entire belief system". In which case, Lycan's view is inconsistent with moderate foundationalism, as that is a view which claims that there are some beliefs which are both non-inferentially justified and which can, by themselves, justify other beliefs.

Lycan's view is, at best, a version of a view that Bonjour calls *weak foundationalism*, according to which non-inferentially justified beliefs exist but they "possess only a very low degree of epistemic justification on their own, a degree of justification insufficient by itself either to satisfy the adequate-justification condition for knowledge or to qualify them as acceptable justifying premises for further beliefs" (Bonjour 1985, p. 28). Bonjour assumes that the commitment of this view to non-inferential justification suffices for it to be a form of foundationalism. However, Poston (2012) provides a strong case against this assumption. And Bonjour himself thinks (1985, p. 29) that this view is perhaps best seen as being a kind of hybrid of foundationalism and coherentism.

For our purposes, the interest of Lycan's position is that it helps us to assess the implications of these two claims about the method of RE: (i) that the input beliefs into this method must already be justified to some degree; and (ii) that this justification is non-inferential, being based directly on the mere fact of the subject's believing or intuiting as they do. What Lycan's view shows us is how one can accept these claims and still offer a version of the method of RE that has strong coherentist features. In particular, one might offer a version of RE on which an initial belief automatically possesses some degree of non-inferential justification, but this degree of justification is tiny, and this belief can only participate in the justification of other beliefs insofar as it is part of a system of beliefs that approximates a state of reflective equilibrium. A view that would have the significant virtue of being able to accommodate the strongly coherentist statements made by Rawls and Goodman,[9] as well as the fact that many coherentists identify their views as being closely related to, or simply versions of, the method of RE, including Lycan himself (1998, p. 212) and others like Elgin (1996, ch. 4) and Sayre-McCord (1996).

5. Objections

The most common kinds of objections made to RE all claim, in different ways, that this method is too weak (Kelly and McGrath 2010). That is, these objections all contend that one could apply this method perfectly to, say, one's initial moral beliefs, and yet the final set of beliefs that one would end up with, or the way in which one formed those beliefs, would still be criticizable in a way that undermines the credentials of RE as being a good method for figuring out what to believe about morality (or knowledge, or logic, etc.). And, typically, these weakness objections identify the reliance of this method on our initial beliefs or

[9] For example, consider the following passage from Rawls (1971, p. 579): "I have not proceeded then as if first principles, or conditions thereon, or definitions either, have special features that permit them a peculiar place in justifying a moral doctrine. They are central elements and devices of theory, but justification rests upon the entire conception and how it fits in with and organizes our considered judgments in reflective equilibrium. As we noted before, justification is a matter of the mutual support of many considerations, of everything fitting together into one coherent whole."

intuitions as being the source of the relevant problems. My aim in what follows is to identify some of the main objections of this kind and to indicate how proponents of the method of RE might reply to them.

5.1 *Objections from Conservatism*

One objection of this kind—that featured in prominent early criticisms of Rawls by Brandt (1979), Hare (1973), and Singer (1974)—is that the method of RE is a disguised form of moral intuitionism and, as such, is open to criticism for being too conservative in the importance it places on our moral theories conforming with our pre-theoretical moral beliefs or intuitions. This conservatism objection is often supported by suggestions that these initial beliefs may stem from untrustworthy sources. For example, Singer (1974, p. 516) writes that "all the particular moral judgments we intuitively make are likely to be derived from discarded religious systems, from warped views of sex and bodily functions, or from customs necessary for the survival of the group in social and economic circumstances that now lie in the distant past".

In response to this conservatism objection, proponents of RE often point to the fact that on this method the contents of one's initial beliefs are not meant to serve as fixed 'data points', to which any acceptable theoretical principles must conform. Rather, wherever there is a conflict between 'intuition' and 'theory' there is always the possibility that the best response to that conflict will be to reject one's initial beliefs. Furthermore, proponents of RE also point out that on the method of WRE there will be the potential for "far more drastic *theory-based* revisions of moral judgments" (Daniels 1979, p. 266).

One way of thinking of this revisability response would be to think of it as saying that, while it is always a cost to a theory if it clashes with one's initial intuitions, this cost can be outweighed by the benefits of adopting that theory. But proponents of the method of RE often have a more radical form of revisability in mind such that the proper set of considered judgements against which our moral theories are to be "checked" (Rawls 1971, p. 51) should only be thought of as the judgements one holds at the *end* of the RE process (see e.g. Daniels 1979 fn. 17; and Scanlon 2003, p. 149).[10] And on this interpretation of the method of RE it may be *no* cost to a theory at all if it clashes with our *initial* moral judgements or beliefs.

5.2 *Objections from Disagreement*

Despite the above revisability points, there is no doubt that our initial beliefs play a significant role in the method of RE. The worry can linger then that there is something objectionable about this fact. One source of concerns about this feature of the method of RE is the thought that different people may have very different initial beliefs and, hence, might reach different equilibria when they apply this method. For example, this thought can lead to the

[10] See also DePaul's (1987) related distinction between 'conservative' versus 'radical' interpretations of the method of RE.

worry that if RE really is our best method for forming beliefs about a given domain then we will be forced into adopting an anti-realist view of that domain.

The issue of whether the moral views of different people would converge after some form of idealized moral inquiry, has often been thought to have implications for moral realism. Rawls himself (1974, p. 9) suggests that if the views of different people would not converge after they applied the method of WRE, then it follows that there are no objective moral truths. It might also seem very plausible that there would be no such convergence, given that there can be substantial cross-cultural differences in our initial moral beliefs (Brandt 1979, p. 22). In which case, one objection that might be made to RE is that it can't be the best method of deciding what to believe about morality because assuming that it is will lead us to some form of anti-realism.

Of course, one may not view this supposed consequence of RE as constituting an objection to this method if one is prepared to abandon moral realism.[11] One might also be sceptical that there is any plausible method of moral inquiry such that everyone who impeccably employed that method would converge on the same moral views. In which case, this supposed consequence could not constitute a unique problem for the method of RE. But, perhaps most importantly, it is simply hard to see how the existence of moral disagreement is meant to support the denial of moral realism (see Enoch 2009 for a detailed discussion of this issue).

There are other disagreement-based objections that might be made to the method of RE. For example, one might object that this method is unjustifiably *arbitrary* in the way it relies on the initial beliefs or intuitions of the person who happens to be employing it, as opposed to the (perhaps conflicting) beliefs of someone else. After all, it would seem to be both egocentric and ethnocentric to adopt a method of inquiry which enjoins one to treat one's own intuitions, or the intuitions of one's culture, as being a reliable guide to the truth, whilst not assigning the same status to the intuitions of other individuals or cultures. As Ichikawa (2014) discusses, Stich (1990) and Weinberg et al. (2001) both seem to appeal to something like this arbitrariness objection when they criticise the method of RE. Stich (1990) supports this objection by appealing to the possibility of cross-cultural disagreement in our epistemic intuitions, and Weinberg et al. (2001) support it by providing experimental evidence that there are actual disagreements of this kind.

One concern with this arbitrariness objection[12] is that it relies on a misunderstanding of the role that initial beliefs or intuitions play in the method of RE. In particular, it seems to rely on something (roughly) like the following two ideas: (i) the initial inputs into this method are propositions of the form *I/we have the intuition that p* which are then treated as evidence for *p*; and (ii) this method does not assign the same evidential role to

[11] Rawls (1974) suggests that the method of WRE does not presuppose the existence of objective moral truths. See Holmgren (1987) for critical discussion of this claim, and Rawls (1980) for related discussion regarding his constructivist approach to moral theory.

[12] The presentation given here of this arbitrariness objection differs from that found in Stich (1990) and Weinberg et al. (2001), as their version focuses on a specific worry about the normative force of conclusions reached by the method of RE. However, I think the kind of worries I have raised will also apply to their version of this objection. It is also important to note that more recent experimentalist critiques of the use of intuitions in philosophy do not rely on this kind of arbitrariness objection (see e.g. Weinberg 2007). See Ichikawa (2014) for an excellent discussion which supports both these points.

propositions of the form *they have the intuition that p*. And if we think of the method of RE in this way then it is easy to understand why someone might view it as being egocentric or ethnocentric.

However, as discussed in §2, the inputs into the method of RE should not be understood in this psychologistic way. On the method of RE one does not begin (say) epistemological inquiry with propositions about one's epistemological intuitions. Rather, one begins inquiry with the contents of one's epistemological beliefs or intuitions, the rough idea being that one can justifiably treat the content of one's belief or intuition that *p* as being provisionally correct directly in virtue of one's believing or intuiting that *p*. As opposed, that is, to treating *p* as true in virtue of it being the conclusion of some inference based on the premise that one believes or intuits that *p*. Once one makes note of this point, the arbitrariness worry appears to lose much of its bite, as it is surely not arbitrary to begin inquiry with the contents of one's own mental states. Indeed, what else could we do?[13]

5.3 *Objections from Error*

The possibility of people reaching conflicting views when they properly employ the method of RE—because of their different initial beliefs—points to another important concern with this method, namely, that there is no guarantee that it will lead us to the truth. For if a method leads one person to believe that *p*, and another to believe that *not-p*, then, obviously, that method does not always lead to the truth. But the mere fact that a method of inquiry may lead to error, even when it has been impeccably applied, is not a good objection to that method. After all, as Kelly and McGrath (2010, p. 326) point out, the same criticism could be applied to the scientific method, as even the scientific method will consistently lead us to falsehoods if we have the misfortune to be in a world where "the empirical evidence that we have to go on is consistently misleading or unrepresentative".

Perhaps the error objection can be reformulated so as to avoid Kelly and McGrath's (henceforth 'K&M') overgeneralization worry. One might try to argue that the problem with RE is that, unlike the scientific method, it is unlikely to lead one to the truth even when it is employed in "normal" conditions. But even if that idea could be made both precise and persuasive it is not clear that it would constitute a good objection. This is because proponents of RE are typically willing to concede that the method is not truth-conducive, whilst denying that this undermines the credentials of the method as a means of acquiring justified beliefs.

[13] More should be said about this issue but for reasons of space these brief remarks will have to suffice here. For related discussion, see Wedgwood (2010) including the following passage (p. 239–40): "It does not seem possible for me currently to form a moral belief *directly* on the basis of *your* moral intuitions. At best, I can only directly base my current formation of a moral belief on my *beliefs* about your moral intuitions. On the other hand, it *is* possible for me currently to form a moral belief directly on the basis of *my own current* moral intuitions. Moreover, it seems that we are disposed to be guided by our moral intuitions towards forming the corresponding moral beliefs: if I currently have a moral intuition, that moral intuition will immediately incline me to accept the corresponding moral belief (unless I have some special reason for doubting that intuition)."

5.4 *Objections from Unreasonable Belief*

K&M (2010) argue that the key problem with RE is not that it is too conservative, or that it might lead us to false beliefs, but rather that it might lead us to beliefs that are unreasonable for us to hold. And if this claim is true it would constitute a powerful objection, given that what is meant to recommend this method is that it is a way to acquire justified beliefs (which presumably cannot be beliefs that are unreasonable for one to hold).[14]

According to K&M, even if the initial inputs into the method of RE have to meet Rawls' constraints on "considered judgments" it will still be possible that someone could impeccably apply this method and yet end up with beliefs that are unreasonable for them to hold. To support this claim, K&M give the example of a subject whose initial considered moral judgements include the judgement that the following proposition is true that I will label 'KILL':

> (KILL) One is morally required to occasionally kill randomly.

As K&M (2010, p. 347) note "there is nothing *incoherent* about the possibility that someone could confidently and stably subscribe to this judgment, even if he or she is aware of all of the non-moral facts, does not stand to gain or lose depending on whether it is true or false, and so on." But then it seems possible that a subject like this could impeccably apply the method of RE and end up with a final set of moral beliefs that still includes this perverse belief that KILL is true. Let us call this hypothetical subject 'Bill'. K&M think it is clear that it would not be reasonable for Bill to hold this perverse belief and, hence, they conclude that impeccably applying the method of RE does not suffice for one to end up with reasonable beliefs.

K&M think the existence of this problem is obscured by the fact that when a proponent of RE illustrates her method she typically proceeds 'from the first person perspective, and speaks of (e.g.) "our" considered judgements; she thus selects one of her own considered judgements that she expects her readers to share' (2010 p. 347). But K&M think that this choice of example raises the worry that the basis for our agreement that the content of this judgment would be a good place to begin inquiry may not be the fact that it is the object of a considered judgement, but rather that it is a proposition which we perceive to possess some positive normative status, like being rationally credible or known to be true. K&M suggest that to decide between these two possibilities we should conduct a certain kind of 'experiment':

> In order to test the claim that it is the fact that the judgment in question is a considered judgment which is doing the work in this context, it is important to consider cases from the third person perspective, in which the starting points of the person pursuing reflective equilibrium are (i) his considered judgments but (ii) *perverse* considered judgments, at least when judged by one's own lights. (That is, judgments which, when judged by one's own lights, are clear cases of nonknowledge, or propositions utterly lacking in rational credibility.) (2010. p. 347)

[14] If this claim is true it would also seem to constitute an objection to DePaul's (1993 and 1998) version of the method of RE on which this method is not a reliable means of acquiring justified beliefs but is a means of acquiring rational beliefs.

And K&M claim that when one performs this experiment (e.g. by considering someone like Bill), one finds that "the idea that the normatively appropriate starting point for a person consists of all and only her considered judgments increasingly loses its appeal" (p. 347).

According to K&M, the moral here is that the method of RE is only defensible if it is reconceived so that the initial beliefs it relies on do have to possess some positive normative status. As an example, K&M consider the possibility of a proponent of RE who rejects (1) in favour of (2):

(1) For any individual, the appropriate starting point from which to pursue wide reflective equilibrium is the class of judgments consisting of all and only her considered judgments.
(2) For any individual, the appropriate starting point from which to pursue wide reflective equilibrium is the class of all and only those judgments that she is justified in holding at that time.

K&M suggest that a proponent of RE who replaces (1) with (2) can avoid the perversity problem. But K&M also close their paper by querying whether the resulting method really deserves the name 'the method of reflective equilibrium'.

What should we make of this perversity objection, K&M's proposed solution to it, and the supposedly negative consequences of adopting that solution? Starting with the objection, K&M are surely right that it is possible for a subject to impeccably apply the method of RE to their initial beliefs and nonetheless end up with beliefs with crazy contents. But could a proponent of the method of RE still resist K&M's further claim that we should thereby reject the method of RE (as it is standardly interpreted)?

One point that is worth noting is that K&M do not provide any arguments in support of their claim that it would not be reasonable for Bill to believe that KILL is true even after applying the method of RE. Presumably, this is because they take this claim to be too obviously true to warrant any such arguments. But I think there are good reasons to be wary of this claim that, in a way, mirrors K&M's own concerns about our agreement that a given considered judgement would be a reasonable place to begin inquiry, when we ourselves endorse that same judgement.

K&M's worry (as I understand it) is that in such cases our agreement might stem from our perception that the content of this judgement has some positive normative status (e.g. being rationally credible or known to be true). In which case, our agreement might not reflect a belief that this proposition would be an appropriate place to begin moral inquiry for anyone who has the considered judgment that it is true. Rather, it might merely reflect a belief that this proposition would be appropriate place to begin moral inquiry for people like ourselves, who are aware that this proposition has this positive normative status, or who stand in some relevant normative relation to that content (like knowing it to be true). .

The roughly analogous worry for K&M concerns their assumption that a perverse proposition like KILL could not be part of an appropriate *end point* for inquiry into the relevant domain. The worry is that any inclination we have to agree with this claim might merely stem from our perception that there are overwhelmingly good reasons for rejecting this proposition. In which case, our agreement might only manifest a belief that this proposition could not be part of an appropriate end point of inquiry for people like ourselves,

who are aware that this proposition has this negative normative status, or who stand in some relevant normative relation to it (e.g. knowing it to be false). As opposed, that is, to reflecting a belief that such judgements could never form an appropriate end point for anyone who inquired into that domain, no matter what initial beliefs they started with.

To support this worry, it might be useful to consider two different ways of thinking of Bill. On the first way we think of Bill's initial set of beliefs as being as close to our own as they could possibly be after adding just this one perverse belief. If we think of Bill in this way, it seems likely that if he impeccably applies the method of RE then this perverse belief should be quickly eliminated. This is because KILL will fail to cohere with all manner of other particular and general moral propositions that Bill initially believes to be true. On the other hand, we might think of Bill as having many other perverse moral beliefs (or, alternatively, we might think of Bill as only having this one perverse belief but then imagine that his confidence in this belief is far higher than his confidence in any of his other beliefs). If we think of Bill in this way then it seems likely that he will still end up with perverse moral beliefs after he applies the method of RE. But note that Bill's initial doxastic situation is now quite radically different from our own. And, once we think of Bill as being this different from ourselves, then I think it becomes far less obvious that it would be unreasonable for him to hold the perverse (from our perspective) beliefs that he might end up with after impeccably applying the method of RE.

For these reasons I think it is far from clear that K&M's perversity objection succeeds.[15] But, for the sake of argument, let us assume that it does. Does K&M's proposed solution to this problem succeed, and are they right to claim that this solution has negative consequences for the method of RE?

Starting with the solution, consider K&M's specific suggestion of replacing (1) with (2). I think it is clear that this proposed solution will fail to block the perversity problem. For example, consider, again, those versions of the method of RE that endorse some form of epistemic or phenomenal conservatism. These are views on which the inputs into the method of RE are meant to be justified beliefs. But it a well-known consequence of both epistemic and phenomenal conservatism, that a subject could be justified in holding crazy beliefs. This is because it seems perfectly coherent that there could be subjects who have perverse seemings or spontaneous beliefs whilst also being unaware of any relevant defeaters. In which case, if either epistemic or phenomenal conservatism is true then a proponent of RE cannot solve the perversity problem by simply replacing (1) with (2).

K&M might reply that this is not a problem with their solution to the perversity objection but rather a problem with conservatism. But I think this issue will generalize much further, indeed to almost any view which endorses the standard assumption that doxastic justification is non-factive. For if justification does not entail truth then it is hard to see how we can rule out the possibility of a subject having justified beliefs with perverse contents like KILL. In which case, it seems that a version of RE which restricts its inputs to justified beliefs will still face the perversity problem.

Perhaps one might try to save K&M's solution to the perversity problem by either defending the controversial thesis that doxastic justification is factive, or by appealing to some

[15] See Lycan (2012, pp. 11–12) for a broadly related defence of his coherentist theory of justified inference against the objection that it will have to sometimes classify crazy beliefs as being justified.

other normative status which is unquestionably factive (like, say, knowledge). But if the supposed moral now of the perversity objection is that the method of RE is only plausible if it is restricted to factive attitudes of some kind, then it seems like this objection is threatening to just collapse back into the objection that the method of RE need not lead us to the truth. But, as we have seen, K&M themselves point out that this kind of objection is problematic and their objection is meant to be independent of the error objection.

Our reflections on the perversity objection suggest that K&M have not provided a compelling case for thinking that the initial inputs into the method of RE have to possess some positive normative status. But, as we have already seen, there are proponents of RE who already accept that the inputs into this method have to possess[16] some (perhaps tiny) degree of non-inferential justification. If only for this reason then, it is worth considering K&M's closing remarks in which they suggest that any such version of the method of RE will face certain negative consequences.

K&M suggest that if one requires that the initial inputs into the method of RE have to be justified beliefs then they thereby lose one of the main supposed virtues of this method, namely, that it allows one to avoid positing some mysterious faculty as the source of our justified beliefs about the relevant domain. This is because one will now need to tell some further story about how our initial beliefs come to be justified. In which case, K&M claim that "the need for a certain kind of traditional epistemological theorizing (with all of its attendant pressures towards postulating non-obvious normative mechanisms, and so on) seems to have re-emerged" (2010, p. 353). Furthermore, K&M suggest that the resulting method may not deserve to be called a version of the method of RE because it will now be "natural to think that the most interesting part of the story concerns not the pursuit of equilibrium itself, but rather what makes it the case that certain starting points are more reasonable than others, and how we manage to recognise or grasp such facts" (2010, p. 354).

Starting with the interest worry, it is hard to see why we should accept this suggestion. The assumption that bringing one's moral beliefs into equilibrium is only part of the story about how one's beliefs get to be justified does not give us any reason to think that whatever else is needed to complete that story will be of greater interest than the equilibrium part of that story. Suppose a proponent of the method of RE grants that the initial beliefs identified at stage 1 must already be justified to some degree, and by something other than their coherence relations to other beliefs. This claim is perfectly consistent with the idea that coherence still plays a very significant and interesting role in justifying our beliefs. For example, as discussed in §4.3, one might hold that this degree of justification is tiny, and that initial belief can participate in the justification of other beliefs if they are part of a system of beliefs that is in a state of reflective equilibrium.

What about K&M's further worry that the explanation of how our initial moral beliefs are justified will have to be a story that appeals to some mysterious faculty of intuition? Again, it is hard to see why we should accept this suggestion. Consider, once more, those versions of the method of RE which endorse some form of epistemic or phenomenal conservatism. Whatever one thinks of such principles, it is by no means obvious that they

[16] Or provide, if we think of these inputs as being intuitions which are then understood as some kind of non-doxastic seeming state (see e.g. Pust 2000 for the suggestion that the inputs into method of RE should be understood in this kind of way).

commit one to some mysterious faculty of moral intuition. Indeed, proponents of these principles often claim that one of their virtues is that they can help to explain how we can acquire certain kinds of justified beliefs without appealing to such mysterious entities.[17]

6. Intuitions and RE

A lot more could be said about objections to the method of RE. However, in closing, I want to briefly comment on a different issue, namely, the relationship between this method and the idea that intuitions play a central role in philosophical inquiry.

The "experimental philosophy" (or "X-Phi") movement has helped to generate a large literature on intuitions and their supposed role in philosophy. Notably, the very first X-Phi paper—Weinberg et al. (2001)—identifies the method of RE as the most familiar example of *Intuition Romanticism*, which is the authors' name for the intuition-based strategy for forming beliefs that is the subject of their experimental critique. The method of RE is also identified as an example of a problematic intuition-based method of inquiry in earlier works that influenced the X-Phi movement (see e.g. Stich 1990 and Cummins 1998), and defenders of the role of intuitions in philosophy have also identified the method of RE as an example of the general position they are seeking to defend (see e.g. Pust 2000, ch. 1).

The method of RE has been strongly linked then with this widespread assumption that intuitions play a central role in philosophical inquiry. But, interestingly, I think the standard characterizations of the supposed role of intuitions in philosophy often diverge in significant ways from the role that our initial beliefs are meant to play in the method of RE.[18] One example concerns the role of intuitive counterexamples in evaluating theories. In the intuitions literature it is widely assumed that a conflict with our pre-theoretical intuitions always counts against a theory, even if that cost can be ultimately outweighed by the benefits of adopting that theory (see e.g. Weatherson 2003, p. 8). On the other hand, as we saw in §5.1, proponents of the method of RE suggest that a conflict with our initial beliefs or intuitions need not count against a theory at all if the content of that intuition is not judged to be true at the end of the RE process.

Another example concerns the assumption that intuitions "serve as a kind of rock bottom in philosophical argumentation" such that "*Intuitive judgments justify, but need no justification*" Cappelen (2012, p. 112). [19] As Cappelen discusses (2012, pp. 118–22), this idea

[17] See e.g. Lycan (1998, pp. 207–15) for related discussion on the relationship between conservatism and the justification of our moral beliefs.

[18] There are also differences in the characterizations offered of the nature of our initial beliefs or intuitions. For example, as mentioned in §3, Rawls (1951) requires that his considered judgements be intuitive. But, as was also noted earlier, Rawls has a very undemanding notion of an intuitive judgement according to which they are simply judgements that are not consciously based on the application of theoretical principles. In which case, Rawls is not committed to the ideas—often found in the intuition literature—that intuitive judgements have a special phenomenology, or that they are based solely on our conceptual competences.

[19] Cappelen (2012) himself rejects this idea as well as the more general one that intuitions play a central role in philosophical inquiry.

is difficult to make precise and can be developed in a number of different ways. But I think it is fair to say that this foundational-type role for intuitions is importantly different from the role that our initial beliefs are meant to play in the method of RE. As we saw in §4.3, some philosophers argue that the method of RE is best interpreted as being committed to our initial beliefs possessing some degree of non-inferential justification. But, as we also saw in §4.3, even if we grant this point, there are good reasons to think that the method of RE should still be interpreted as holding that our initial beliefs can only participate in the justification of other beliefs insofar as they are part of a belief system that has been brought into a state of RE. This is not a view, then, on which our initial beliefs can justify without themselves needing justification.

Obviously, a lot more should be said about these two examples and the relationship between the method of RE and standard characterizations of the role of intuitions in philosophy. But I think these brief considerations suggest that this relationship may be more complex and interesting than it is usually assumed to be.

References

BonJour, L. (1985). *The Structure of Empirical Knowledge*. Cambridge, MA: Harvard University Press.

Brandt, R. (1979). *A Theory of the Good and the Right*. Oxford: Oxford University Press.

Cappelen, H. (2012). *Philosophy Without Intuitions*. Oxford: Oxford University Press.

Cummins, R. (1998). "Reflection on Reflective Equilibrium". In DePaul and Ramsey (eds.) *Rethinking Intuition*. Lanham, MD: Rowman and Littlefield: 113–27.

Daniels, N. (1979). "Wide Reflective Equilibrium and Theory Acceptance in Ethics". *Journal of Philosophy* 76 (5): 256–82.

DePaul, M. (1987). "Two Conceptions of Coherence Methods in Ethics". *Mind* 96: 463–81.

DePaul, M. (1993). *Balance and Refinement: Beyond Coherence Methods of Moral Inquiry*. London: Routledge.

DePaul, M. (1998). "Why Bother With Reflective Equilibrium?" In Michael DePaul and William Ramsey (eds.) *Rethinking Intuition*. Lanham, MD: Rowman and Littlfield: 293–309.

Ebertz, Roger P. (1993). "Is Reflective Equilibrium a Coherentist Model?" *Canadian Journal of Philosophy* 23 (2): 193–214.

Elgin, C. (1996). *Considered Judgment*. Princeton, New Jersey: Princeton University Press.

Enoch, D. (2009). "How is Moral Disagreement a Problem for Realism?" *Journal of Ethics* 13 (1): 15–50.

Foot, P. (1967). "The Problem of Abortion and the Doctrine of Double Effect". *Oxford Review* 5:5–15.

Goodman, N. (1954). "The New Riddle of Induction". In his *Fact, Fiction, and Forecast*. Cambridge MA: Harvard University Press: 59–83.

Hare, R. M. (1973). "Rawls' Theory of Justice". *Philosophical Quarterly* 23: 241–52.

Holmgren, M. (1987). "Wide Reflective Equilibrium and Objective Moral Truth". *Metaphilosophy* 18 (2): 108–24.

Holmgren, M. (1989). "The Wide and Narrow of Reflective Equilibrium". *Canadian Journal of Philosophy* 19 (1): 43–60.

Ichikawa, J. (2014) "Who Needs Intuitions? Two Experimentalist Critiques". In Anthony Booth and Darrell Rowbottom (eds.), *Intuitions*, Oxford: Oxford University Press, 232–56.
Kelly, T. and McGrath, S. (2010) "Is Reflective Equilibrium Enough?" *Philosophical Perspectives* 24 (1): 325–59.
Lewis, D. (1983). *Philosophical Papers, Volume I*. Oxford: Oxford University Press.
Lycan, W. G. (1988). *Judgment and Justification*. New York: Cambridge University Press.
Lycan, W. G. (2012). "Explanationist Rebuttals (Coherentism Defended Again)". *Southern Journal of Philosophy* 50 (1): 5–20.
McMahan, J. (2000). "Moral Intuition". In Hugh LaFollete (ed.) *The Blackwell Guide to Ethical Theory*. Chichester: Blackwell: 92–110.
Poston, T. (2012). "Basic Reasons and First Philosophy: A Coherentist View of Reasons". *Southern Journal of Philosophy* 50 (1): 75–93.
Pust, J. (2000). *Intuitions as Evidence*. New York: Routledge.
Rawls, J. (1951). "Outline of a Decision Procedure for Ethics". *Philosophical Review* 60: 2:177–97. Reprinted in Rawls (1999): 1–19.
Rawls, J. (1971). *A Theory of Justice*, 2nd edition 1999. Cambridge MA: Harvard University Press.
Rawls, J. (1974). 'The Independence of Moral Theory'. *Proceedings and Addresses of the American Philosophical Association* 47: 5–22. Reprinted in Rawls (1999): 286–302. Page references are to the reprinted version.
Rawls, J. (1980). "Kantian Constructivism in Moral Theory". *Journal of Philosophy* 77: 515–72. Reprinted in Rawls (1999): 303–58. Page references are to the reprinted version.
Rawls, J. (1999). *Collected Papers*, Sam Freeman, (ed.). Cambridge MA: Harvard University Press.
Rawls, J. (2001). *Justice as Fairness*. Cambridge MA: Harvard University Press.
Sayre-McCord, G. (1996). "Coherentist Epistemology and Moral Theory". In Sinnott-Armstrong, W. and Timmons, M. (eds.), *Moral Knowledge? New Readings in Moral Epistemology*. New York: Oxford University Press: 137–89.
Scanlon, T. M. (2003). "Rawls on Justification". In Samuel Freeman (ed.), *The Cambridge Companion to Rawls*. New York: Cambridge University Press: 139–67.
Singer, P. (1974). "Sidgwick and Reflective Equilibrium". *The Monist* 58 (3): 490–517.
Stich, S. (1990). *The Fragmentation of Reason*. Cambridge MA: MIT Press.
Thomson, J. J. (1976). "Killing, Letting Die, and the Trolley Problem". *The Monist* 59 (2): 204–17.
Weatherson, B. (2003). "What Good Are Counterexamples?" *Philosophical Studies* 115 (1): 1–31.
Wedgwood, R. (2010). "The Moral Evil Demons". In Richard Feldman and Ted Warfield (eds.), *Disagreement*. Oxford: Oxford University Press: 216–46.
Weinberg, J. (2007). "How to Challenge Intuitions Empirically Without Risking Skepticism". *Midwest Studies in Philosophy* 31 (1): 318–43.
Weinberg, J., Nichols, S., and Stich, S. (2001). "Normativity and Epistemic Intuitions". *Philosophical Topics*, 29 (1–2): 429–60.
Willimason, T. (2007). *The Philosophy of Philosophy*. Oxford: Blackwell Publishing

CHAPTER 13

ANALYTIC–SYNTHETIC AND A PRIORI–A POSTERIORI HISTORY

BRIAN WEATHERSON

1. Introduction

It's easy to give a rough gloss of the notions of analyticity and a priority:

- Something is an analytic truth iff it is true in virtue of its meaning.
- Something is an a priori truth iff it is knowably true without justification by experience.

And this yields us two distinctions, between analytic truths and synthetic truths (i.e. every truth that isn't analytic), and between a priori truths and a posteriori truths (i.e. every truth that isn't a priori). But fleshing out these distinctions takes some work, as we'll see. Let's start by a quick historical survey of work on the two distinctions, their relationship to each other, and their relationship to the necessary/contingent distinction. There are four important stops in the history.

The distinction between "analytic" and "synthetic" traces back to Kant. He thought that both distinctions in our title (analytic/synthetic and a priori/a posteriori) were real, and that they were not the same distinction. In particular, he held that most interesting philosophical and mathematical claims were synthetic a priori. This is because he (at least most of the time) worked with a fairly narrow notion of analyticity. A subject–predicate sentence *A is B* is analytic if "the predicate *B* belongs to the subject *A* as something that is (covertly) contained in this concept *A*" (Kant 1999, 6). But there can be plenty of a priori truths that do not fall under this narrow category.

Let's look at one example of current importance. Consider the claim *Whatever is known is true*. It is at least plausible that that is a priori: that we don't need to look into the world to know that knowledge implies truth. But is it analytic for Kant? It is iff there is an analysis

of knowledge, and the analysis is of the form *S knows that p iff p is true, and X*, for some value of *X*. Most epistemologists nowadays would reject the idea that there is an analysis of knowledge. But even among those who hold out hope for such an analysis, an argument by Linda Zagzebski (1994) has convinced most people that the analysis cannot be of this form. In particular, Zagzebski argued that if there is such an analysis, *X* must entail that *p* is true, eliminating the need for this clause. So *Whatever is known is true* will turn out, by Kantian standards, to be synthetic, even if it is a priori.

Although Kant did not think the two distinctions we are focussing on are equivalent, most scholars take him to have thought all and only necessary truths are a priori knowable. (Though see Strang (2011) for a dissent.) This will be a common theme throughout much of the history.

Our second stop on the history is logical positivism, most clearly represented in English by Ayer (1936). The positivists thought that all three distinctions were in a fairly deep sense equivalent. In particular, they thought all three were very close to the distinction between theorems and non-theorems of logic. The positivists were, self-consciously, building on the tradition of British empiricism. But unlike some empiricists, they didn't want to insist on an empirical basis for logical and mathematical knowledge.[1] The solution was to build on the logicism about mathematics developed by Frege and Russell.

Borrowing a term from Paul Boghossian (1996), let's say that a sentence is *Frege-analytic* iff it can be converted to a logical truth by the substitutions of synonyms. The positivists thought that all a priori, necessary, and analytic truths were the Frege-analytic truths. Without logicism, this would be wildly implausible, since mathematical truths would be an exception. But Frege and Russell had done enough to make that possibility less worrying.

The positivists' view has some epistemological attractiveness. How can we know things without having empirical input? And how can we know that some things are true, not just in this world, but in all worlds? Well, say the positivists, by knowing the language (which we learn empirically) we learn what sentences are related by the substitution of synonyms. And then the puzzles about knowledge of analytic or necessary truths just reduce to puzzles about the epistemology of logic.

Our third major stop then is with Quine, who questioned both steps of this attempted explanation. First, Quine (1936) noted that the story still needs an epistemology of logic. The obvious expansion of the story told so far won't work. It can't just be by learning the meanings of the logical connectives that we come to learn which the logical truths are. That's because we need to be able to derive the consequences of those meanings, and for that we need logic.

But second, Quine (1951) argued that we have no independent way to make sense of the notion of synonymy that is at the heart of Frege-analyticity. This is the most famous part of Quine's attack on the empiricists' epistemology of logic and mathematics, but it isn't the strongest part of it. Indeed, the argument in "Two Dogmas" is both strange and self-undermining.[2] Quine's primary complaint in that paper about the notions of analyticity, synonymy, and meaning is that the only way we have of understanding these notions is in

[1] Actually, the history of what pre-positivist empiricists believed about mathematics is a little more complicated than the standard story. See Whitmore (1945) for some details.

[2] The next few sentences follow the arguments of Sober (2000) fairly closely.

terms of the others. But that would only be a problem if we thought we needed to understand them in terms of something else. Arguably we need not: the notions could be theoretical primitives. And especially if one is a confirmation holist (and part of the point of "Two Dogmas" is to defend holism) we shouldn't worry about circularity cropping up near the core of our epistemology. Relatedly, a naturalist like Quine shouldn't care about whether we can give a definition of terms like "meaning", but rather about whether it is a useful concept in a science like linguistics or cognitive science.

To avoid attributing an incoherent position to Quine, we should interpret the argument of "Two Dogmas" as part of his larger argument against the appeal to meanings and analyticities.[3] Quine's larger point, as defended in Quine (1960), was that meanings were unnecessary scientific postulates. He thinks that we simply don't need them to explain all the facts about cognition and communication that need explaining. Now it isn't clear how many people will share Quine's view that meanings are unnecessary for these sciences, since without his behaviourism the attempt to do without meanings looks unsuccessful. But the larger point is that Quine isn't simply relying on an argument from the irreducibility of analyticity to a dismissal of the analytic/synthetic distinction.

The last stop on our history tour is semantic externalism. The externalists complicated the above story in two overlapping ways. First, they developed convincing arguments that necessity and a priority were dissociable. The most compelling of these arguments were the examples of necessary a posteriori truths, such as *Water contains oxygen*. No matter what surface characteristics or functional roles a substance might play, if it does not contain oxygen, it could not be water.

Second, they showed that the pre-theoretical notion of meaning, which had seemed good enough for much prior philosophical theorizing, contained a number of distinct ideas. Here is how Gillian Russell (whose writings I've leaned heavily on in this introduction) puts it,

> In three astonishingly influential pieces of philosophical writing, Hilary Putnam (1973) argued that meaning couldn't be both what a speaker grasped and what determined extension. Kaplan (1989) argued that what determines extension (character) and what got contributed to what a sentence said (content) came apart in the cases of indexicals and demonstratives, and Kripke (1980) argued that what determined the extension of a name or natural kind term need not be known in order for a speaker to understand the expression, nor was it what was contributed to the proposition expressed by a sentence containing one. Each was suggesting that the roles attributed to a single thing—the expression's meaning—in the [pre-theoretical] picture, can be played by distinct things.
>
> (Russell 2008, x)

From this point on, when we talk about truth in virtue of meaning, we have to clarify which aspect of 'meaning' we mean. With that in mind, let's turn to the questions that have been raised about the distinctions.

[3] This paragraph follows closely the discussion of Quine in Russell (2008).

2. Five Questions

To focus our discussion, let's start with five questions we could ask about either the purported distinction between analytic and synthetic, or between a priori and a posteriori.

1. Is there a sensible distinction here?
2. Are there truths on either side of the line?
3. Does the distinction track something of independent significance?
4. Do all distinctively philosophical truths fall on one side of the line?
5. Is the distinction relevant to philosophy?

The questions are obviously not independent: a negative answer to the first suggests that we better not offer a positive answer to any of the rest, for example. But there are more degrees of freedom here than might immediately be apparent.

A negative answer to the second question, for instance, need not imply a negative answer to the first. If one held, with Phillip Kitcher (1980) that a priori warrant is by its nature indefeasible, and as a matter of fact no warrants are indefeasible, then one would think the a priori/a posteriori distinction is sensible, but in fact everything falls on one side of it.

With respect to the a priori/a posteriori distinction, I'll argue that while the answer to question 4 is clearly negative, the answer to question 5 is positive. The a priori/a posteriori distinction may be relevant to philosophy even if it isn't relevant to, for example, demarcating philosophy from non-philosophy. Alternatively, a positive answer to question 5 may follow from a negative answer to one of the earlier questions. (Williamson (2013) suggests, but ultimately I think does not endorse, the view suggested in the following sentences.) If we learned that all knowledge was a posteriori, that is, that all knowledge depended in an epistemologically significant way on experience, that would be epistemologically interesting. So the distinction might have a valuable role in articulating, and perhaps defending, a key philosophical insight, even if all the actual cases fall on one side of the distinction.

The point of raising these questions at the start is to ward off a possible confusion that can easily arise when discussing distinctions. It is common to hear about 'attacks' on a distinction, or 'scepticism' about a distinction, but a moment's reflection shows that it isn't clear what this comes to. I think that most of the 'attacks' on either of our two distinctions are arguments for a negative answer to one of these five questions. (We'll see some instances of this as we go through the chapter.) But different attackers may argue for different negative answers, and different defenders defend different positive answers. So it is, I think, helpful to have these distinct questions in mind before we begin.

3. The Traditional Notion of the A Priori

The traditional notion of the a priori makes the best sense, I think, if you start with the following three assumptions.

1. There is a notion of justification that is distinct from, but a constituent of, knowledge.
2. Whether a belief is justified, in this sense, depends just on the evidence the believer has.
3. Evidence about the external world consists solely of perceptual experiences.

From 2 and 3 we get the idea that there could be some beliefs whose justification does not depend on any perceptual experience, that is, beliefs that are justified by a null set of perceptual experiences. These are the beliefs that are justified first, that is, a priori. Then by 1 we can say that these beliefs satisfy a part, possibly a large part, of the conditions for being knowledge. And this is the a priori knowledge.

The problem, as will probably be clear to most readers, is that all three of the assumptions I started with are contentious. As noted in the introduction, Linda Zagzebski (1994) has shown that there cannot be any non-factive notion of justification that is a constituent of knowledge. Timothy Williamson (2000, ch. 8) has argued convincingly against the phenomenal account of evidence: our evidence consists of facts about the world, not just facts about our experience. Point 2 is less clearly mistaken, but is still far from obvious. (See Conee and Feldman (2004) for a long defence of point 2, as well as discussion of several problems with it.)

Once we drop the three ideas though, or even just the first and third, what could be left to say about the a priori? A natural first move is to think about what explains a person's knowledge, rather than what constitutes it. On the classical picture I just sketched, Bob's knowledge that there are tigers nearby might be constituted by his experience of hearing tiger-like growls. On that picture, having that experience is (partially) constitutive of being justified in believing that there are tigers nearby, and that justification is (partially) constitutive of his knowing there are tigers nearby, and it is these constitutive connections that make his knowledge a posteriori. We don't need to make assumptions that are nearly so strong to conclude that the experience partially explains his knowledge. The experience could (partially) explain why he is justified, without being any part of the justification, and the justification could (partially) explain why he knows, without being any part of the knowledge.

But there's a problem with this move too. I know that all tigers are tigers. On a standard view about the a priori, this will be a piece of a priori knowledge. But to explain why I have that knowledge, you have to appeal to some experiences I have had. After all, with no experiences, I would not be able to think about tigers. So maybe nothing will end up a priori.

There's another usual response here. The experiences I have enable me to think about tigers, without doing anything to justify my belief that all tigers are tigers. So maybe a priori knowledge is that knowledge where experiences do not play any justificatory role, although they may play an enabling role.

That distinction between justifying and enabling will do a lot of work in what follows, so it is worth pausing over it. Perhaps we can say a bit more precisely what it means. An experience is a mere enabler if it explains why a person knows that *p*, but not in virtue of explaining how it is they can believe that *p*. I think something like that is plausibly true, but it still makes a rather large epistemological assumption, namely that justification is explanatorily prior to knowledge. That's something that will be rejected by those who accept the 'knowledge first' epistemology of Williamson (2000).

If you don't accept that justification is explanatorily prior to knowledge, this route at least to articulating the difference between experiences that enable, and experiences that justify, is closed off. And perhaps the enabling/justifying distinction is too obscure to do much work. That's what Williamson has recently argued, and in the next section we'll look at his argument.

4. A Priori Knowledge and Practical Skills

The a priori/a posteriori distinction, on the best way of freeing it from outdated epistemological assumptions, relies on the idea we can make sense of the idea that some experiences are necessary for knowledge because they enable that knowledge, rather than that they justify that knowledge. Timothy Williamson (2013) has argued that this distinction is too unclear to be useful, and as a result the a priori/a posteriori distinction cannot do the work epistemologists need.

I find the example Williamson uses, involving Norman and *Who's Who* (Williamson 2013, 295), rather unintuitive, so I'll substitute a different example that I think makes the same point. Diane is a great basketball player. One of her great skills is being able to anticipate the moves a defence will make, and responding with a move that will maximize her team's chance of scoring. This is a skill she's honed through years of practice and competition. And her most common manifestation of it comes in game situations, when she sees an opposing defender and realizes what move will maximize her team's expected points. But she can also manifest this skill 'off-line', when she considers conditional questions of the form *If the opposing team were to do this, what should I do?*.

Williamson notes that in some such cases, these questions will be solved through the use of imagination, which is surely right at least for some sense of 'imagination'. And in these cases, there won't be any particular experience that justifies the answer. Yet Diane can acquire knowledge by these acts of the imagination. Is the knowledge she gets a priori or a posteriori? Williamson thinks there is no good answer to this question, since Diane's experiences play a role in honing her skills that goes beyond the enabling role, but this role is very different from the role experiences play in classical examples where we can point to a particular experience that justifies the answer.

Now it might seem that there's a simple move to make here. Diane's knowledge is obviously a posteriori because it is explained by her years of experience playing basketball. It is a case of (massive) overdetermination, but that doesn't mean the experiences collectively are not an essential part of the explanation. Williamson's response is that if we go down this route, some paradigmatic instances of a priori reasoning will turn out to be a posteriori. For instance, our ability to engage in logical reasoning might turn out to be dependent on our ability to track the identity of objects (or even just terms) across time. In general, this kind of response threatens to drive the a priori out of philosophy altogether.[4]

[4] Note this argument is distinct from the argument Williamson (2007) makes about the role of knowledge of counterfactuals in philosophical reasoning, and its implications for the a priori status of philosophical knowledge. We'll return to that argument, and the response by Ichikawa and Jarvis (2009). The key point is that this argument only turns on the idea that philosophical reasoning might rest on empirically acquired and honed skills.

This is not, I hasten to note, Williamson's conclusion. Williamson thinks that some of our logical knowledge is a priori, and Diane's knowledge is a posteriori, but the salient explanations of how those pieces of knowledge are obtained and sustained are similar in epistemologically salient respects. So he concludes the a priori/a posteriori distinction does not track anything of epistemological significance.

5. Innate Knowledge

In section 4.1 we considered an argument that there is much less a priori knowledge than we usually assume. In this section we'll look at an argument, tracing back to work by John Hawthorne (2007) that there is a lot more a priori knowledge than we ever thought, in principle a lot lot more.

Recall that we've argued, on pain of losing all a priori knowledge, that we must understand a priori knowledge as knowledge that is in some sense prior to experience, not knowledge that is independent of experience. So now consider beliefs that really are prior to experience, namely innate beliefs. There is a lot of evidence that neonates have differential reaction to (right way up) human faces than they do to other objects. (See Chien (2011) and Heron-Delaney, Wirth, and Pascalis (2011) for some recent studies on this and citations to many more.) It is natural to explain this by positing an internal representation in the neonate of the structure of human faces; that is, a belief about how human faces are structured. Since these beliefs are true, and are in a good sense held because they are true, they seem to amount to knowledge. Yet they are clearly not grounded in, or explained by, the experiences of the neonate. So they look like a priori knowledge.

This is obviously very different to the standard conception of what is a priori knowledge. As we noted in the introduction, and will expand on in the section 13.6, there is a lot of interest in the possibility of a priori knowledge of contingent truths. But even the most enthusiastic supporters of the a priori don't think we have a priori knowledge of facial structure of conspecifics.

The problem is actually worse than this. We don't normally focus on what is actually known a priori, but what is a priori knowable. The reason for this is fairly simple. Most of us cannot know complicated enough multiplications without the aid of empirical evidence.[5] But this doesn't compromise the idea that mathematical truths are in a deep sense a priori. That's because one could, in principle, know them a priori, even if creatures with small brains or limited skills need assistance from their perceptions.

But if that's right, then we can imagine creatures with all sorts of different innate beliefs. Indeed, for any law about the world, we can imagine a creature who innately believes that law to hold, and whose belief has the right kind of evolutionary explanation for it to count as knowledge. (Why the restriction to laws? Well, beyond that there might be issues about

[5] I think that when one carries out multiplication by hand, using the techniques taught at school, the marks on the paper play a justifying and not an enabling role. But arguing for that would be beyond the scope of this entry. It should be less controversial that multiplications carried out by machine give us a posteriori knowledge of the answer.

whether the innate beliefs are accidentally true. In any case, I'm not claiming that only laws could be known a priori this way.)

There isn't any obvious way out here. I think the best thing to do is to say that when we say that something is a priori knowable, we have to mean that it is a priori knowable for creatures like us. That rules out the possibility of having all the laws be a priori, but at the cost of making some arithmetic truths a posteriori.

The arguments by Williamson and Hawthorne I've discussed in sections 13.3 and 13.4 challenge the utility of the traditional notion of the a priori. But for the rest of this chapter I'll set them aside, and discuss what ways we might modify, or use, the traditional notion, should we find responses to these challenges.

6. Substantive A Priori

As we noted in the introduction, a common thread through much of the history of this topic was a belief in a close relationship between a priority and necessity. Most writers take Kant to have treated them as co-extensive notions, the positivists thought they were identical, and Quine took them to suffer from similar defects. It is only with the externalists that we see a gap appearing between the two.

Even once the externalists appear, the gap is not as wide as it may be for two interlocking reasons. The first is that the argument from externalism to the existence of the necessary a posteriori is clearer than the argument from externalism to the existence of the contingent a priori. The second is that externalism may only give a "shallow" distinction between necessity and a priority. Let's take these in turn.

Assuming externalism, we can identify examples of the necessary a posteriori by using familiar natural kind terms. To take a famous example, it is necessary and a posteriori that water is H_2O. It is harder to even identify the contingent a priori. The rough idea is clear enough. We take the characteristics by which ordinary language users identify water, and say it is a priori that water has those characteristics. But what are those characteristics? Water is the stuff which falls from the sky, fills the rivers, lakes, and oceans, and so on. Is any one of these a priori? Not really. It could turn out that nothing had all these properties. (Is it really water in the oceans anyway, or is salt water a different substance?) So we could introduce a new term, Chalmers (1996) suggests "watery" for the long disjunction of conjunctions of properties that a substance must have if we are to identify it as water. Perhaps we come up with a list such that *Water is watery* will be a priori, though since H_2O need not have been watery, it will be contingent. Note we will, at the least, have to introduce new vocabulary to identify this kind of contingent a priority.

The other worry is that the gap opened up here is "shallow" in the sense of Gareth Evans (1979). Given the way things turned out, water must be H_2O. But in some intuitive sense, things could have turned out differently. (I'm taking the helpful locution "could have turned out" from Yablo (2002).) It could have turned out that the stuff in the rivers, oceans, etc. was XYZ. So while it is necessary that water is H_2O, it could have turned out that this not only wasn't necessary, it wasn't even true. If that all sounds plausible to you, you may well think that the a priori truths are all and only those truths which couldn't have turned

out to be false. This way of thinking is behind the important two-dimensionalist approach. Important works in this tradition, as well as Evans (1979), include Davies and Humberstone (1980), Chalmers (1996), and Jackson (1998).

Now there are significant challenges facing two-dimensionalists, several of which are set out in Block and Stalnaker (1999). But my sense is that several of these challenges are very similar to the challenges facing anyone trying to get an argument from semantic externalism to the contingent a priori. If we could say more clearly what it is for something to be watery, it would be easier to say whether a particular world is one where water turned out to be XYZ. So I suspect if externalism gives us a reason to believe in the contingent a priori, it will be a fairly shallow distinction. (This doesn't extend to the argument from externalism to the necessary a posteriori: we don't need to shore up two-dimensionalism to say that *Water contains oxygen* is necessary a posteriori.)

But that's not the only way that a priority might outrun necessity. In recent years there has been a surge of interest in the idea that we can know a priori various anti-sceptical propositions. This idea was advanced in detail by John Hawthorne (2002), and then suggested as a way out of sceptical problems by Roger White (2006) and Brian Weatherson (2005).

Recently, Stewart Cohen (2010) and Sinan Dogramaci (2010) have suggested that (assuming inductive scepticism is false), we can use ampliative inferential steps in suppositional reasoning. That is, if it is possible to inductively infer B when we know A, it is possible to infer B on the supposition that A, and go on to infer the material conditional $A \supset B$. That conditional will be contingent if the inference was ampliative, but since we've discharged the only supposition we used, it could be a priori in a good sense. I have doubts about this route to the contingent a priori (Weatherson 2012), but I think the general idea is plausible.

To make things more concrete, consider "bubble worlds". A bubble world consists of a person, and the space immediately around them. If you think that evidence supervenes on sensory irritation, then you have a duplicate in a bubble world who has the same evidence as you.[6] But you're not in a bubble world, and you know it. There's a well-known probabilistic argument that your evidence can't be grounds for ruling out possibilities that entail you have just that evidence.[7] So your evidence doesn't rule out that you're in a bubble world. But you know you're not. Hence that knowledge is a priori. So *I'm not in a bubble world* might be contingent a priori.[8] Moreover, it's not a "shallow" contingency. It could have turned out that you were in a bubble world. Indeed, with some more evidence you might even know this.

[6] If you prefer a wider conception of evidence, so that for instance two people who are looking at distinct duplicates have distinct evidence, just make the bubble a little bigger, and this argument will still go through.

[7] See White (2006). David Jehle and I have argued that this argument uses distinctively classical logical principles in a way that might be problematic (Jehle and Weatherson 2012). And I've argued that even slight weakenings of the assumptions about how to update credences make the argument fail (Weatherson 2007).

[8] I'm assuming here that the semantic response to scepticism, as defended by Hillary Putnam (1981), doesn't work for ruling out bubble worlds. Defending this assumption would take us too far from the current topic.

I'm not going to defend here the claim that it's contingent a priori that you're not in a bubble world. Indeed, I don't even believe the probabilistic argument for that conclusion that I just referenced. But it is worth noting this trend towards taking seriously the possibility of substantive a priori knowledge.

7. The A Priori in Philosophy

I've argued so far that the best sense we can make of the a priori allows for a lot of a priori knowledge. Once we realize that a priori knowledge, like any other kind of knowledge, is defeasible, and fallible, it seems possible that an agent could have a lot of foundational knowledge of contingent matters. And that foundational knowledge does seem to be a priori. Of course, such an agent would not be very much like us, so there is still a question of what agents like us could know a priori. And it might seem that the class of such pieces of knowledge might be relatively small and interesting.

In particular, one might think that philosophical knowledge, or at least some interesting part of philosophical knowledge, might be a priori. Herman Cappelen (2012) notes that a wide range of philosophers, with very different commitments, end up with the view that the a priori has a distinctive role to play in philosophy. (See, especially, chapters 1 and 6 of that book.) But, as Cappelen also shows, these philosophers are mistaken: outside perhaps of philosophical logic, the a priori doesn't have a particularly special role to play in philosophical inquiry. We can see this, I think, by working through one recent debate.

Timothy Williamson (2007) noted that philosophical thought experiments are almost always incomplete. The text of an example doesn't guarantee that conclusions that are usually drawn from it. To use his example, to guarantee that the subject in one of Gettier's examples has justified beliefs (that don't amount to knowledge), we have to suppose that there are no defeaters in the vicinity, but that isn't stated in the example. Williamson's solution to this is that we should read the example as a certain kind of counterfactual. What we know, Williamson argues, is that in the nearest world where the example was instantiated, the subject would have justified true belief without knowledge. (I'm ignoring here some complications involving names, and donkey anaphora, that are not relevant to this debate.)

Jonathan Ichikawa and Benjamin Jarvis (2009) object that this makes philosophical knowledge a posteriori. We have to know what the world is like to know what the nearest world in which the Gettier case is instantiated is like. Ichikawa and Jarvis reject this because they want to defend a "traditional" conception of thought experiments on which they provide a priori knowledge. I'm not convinced that this really is part of philosophical tradition; it seems to me the thought experiments in Hobbes, Hume, Mill, and many others in the canon rely on empirical knowledge. But I won't press that point here. If one does want to avoid empirical knowledge coming in via the route Williamson suggests, Ichikawa and Jarvis develop a nice way of doing so.

They say that thought experiments are little fictions. We need some empirical knowledge to interpret the fiction. But, they insist, once we are given the fiction, it is a priori that the fiction is possible, and that in it the subject has a justified true belief without knowledge. And it is a priori that these propositions entail that it is not necessary that all justified true

beliefs amount to knowledge. The resulting picture is that some a posteriori knowledge is necessary in interpreting the thought experiment, but once this interpretation is done, the important philosophical conclusion can be drawn a priori.

But let's try and generalize this to other thought experiments. Start, for example, with the famous violinist described by Thomson (1971). That violinist plays a key role in an argument whose conclusion is that abortion is often morally permissible. And one premise of the argument is something about an imagined violinist. Williamson will say that premise is an a posteriori counterfactual proposition. Ichikawa and Jarvis will say it is an a priori proposition about what's true in a fiction. Perhaps there's another premise about the possibility of the example, and maybe that's a priori too. But those premises don't come close to supporting Thomson's conclusion. We need another premise about the analogy between the violinist and a woman contemplating having an abortion to get Thomson's conclusion. Any such premise will not be a priori, unless some detailed facts about human biology are a priori. It's a little tricky, but it isn't obvious the soundness of the abductive inference from those premises to Thomson's conclusion is a priori either. (See Pargetter and Bigelow (1997) for some discussion of this point.)

I think Thomson's example is more typical of philosophical reasoning than Gettier's. We don't just use thought experiments to dismiss theories, like the JTB theory of knowledge. We also use them to defend philosophical conclusions, such as the permissibility of abortion. And in general inferences from a thought experiment to the truth of a theory will involve some a posteriori steps. So even if Ichikawa and Jarvis are right that we can know a lot about thought experiments a priori, it won't follow that in general, philosophical knowledge derived from thought experiments is a priori. Get away from special cases where the facts about the thought experiment entail the philosophically interesting result, and this should be reasonably clear.

That doesn't mean that there's no use for the a priori in philosophy. It might be a very helpful concept to use in argument, even if it isn't true that our conclusions are generally a priori. I'll illustrate with one example from my own work. One way to support the sceptical intuition that we don't know we aren't brains in vats is to ask, how could one possibly know that? Rhetorical questions are not arguments, the received wisdom of undergraduates notwithstanding, so the sceptic needs to find some way to extract argumentative force from that question. An attractive option is argue that *these* are all the ways to know something, and you can't know you're not a brain in a vat any of *these* ways. Such an argument typically runs into problems at the first step: arguing that one has exhausted all possible ways of getting knowledge is not easy.

Hume (1978) had the best idea for how to overcome this step. Don't list the ways someone can know something: use some property of knowledge-gathering methods to partition the methods. Then argue that in no cell of the partition can one find a method that allows knowledge of the undesired kind. If the partition just consists of the presence or absence of some property, you're guaranteed at least to have covered the field. I've argued (Weatherson 2007) that you get an interesting argument by letting the property in question be *is an a priori method*. By "interesting" I certainly don't mean sound. (And nor do I insist that Hume equated interesting sceptical arguments with sound ones.) But I think you get epistemological insight by thinking about whether knowledge that we're not brains in vats could be a priori, or could be a posteriori. You don't have to think that philosophical

conclusions themselves are a priori to think this could be a useful philosophical approach. That last point is probably obvious. We could have developed the sceptical argument by asking whether knowledge of nonenvattedness is innate or acquired. But suggesting that's an interesting argument wouldn't imply the very traditional view that philosophical knowledge is typically innate.

8. Metaphysical Accounts of Analyticity

Let's turn now from the a priori to the analytic. As we noted at the start, the traditional notion is that a sentence is analytic iff it is true in virtue of meaning. And, as we saw at the end of the introduction, this notion is complicated by the fact that traditional theories of meaning conflated several things that should be kept separate.

Paul Boghossian (1996) makes a distinction that has been highly influential between metaphysical and epistemological conceptions of analyticity, and I will follow many contemporary writers in splitting the topic up in this way. The metaphysical conception is the one most continuous with the traditional notion of analyticity, and also the one least popular with contemporary theorists, so we will start with that. It is the notion that some sentences are true merely in virtue of their meaning.

Boghossian, following Quine, argues that this notion is either nonsensical or trivial. Consider a simple example of a putatively analytic truth, say *Everything is self-identical*. Why is this true? In part, because it means that everything is self-identical. But that can't be what we mean to say that it is analytic. *Paris is beautiful* is true in part because it means that Paris is beautiful, but that doesn't make the sentence an analytic truth. What we need is that this is the only thing needed for the sentence to be true. And that isn't the case for either sentence. *Everything is self-identical* is true because of what it means and the fact that everything is indeed self-identical, and *Paris is beautiful* is true because of what it means and the fact that Paris is indeed beautiful. We haven't yet found a difference between the two.

It might be easy to see a response here. Start with a less discriminating treatment of truth makers than I supposed in the previous paragraph. Say that a sentence is true in virtue of what it means, and the way the world is. So both of our examples are true in virtue of their meaning and the way this world is. But for *Everything is self-identical*, it doesn't matter how the world actually is, any way it could have been would have made the sentence true. The contribution of the world is like the contribution of the 5 in *What is 0 times 5?* You need a second number there, or the question doesn't make sense, but it doesn't matter which. In some good sense, the 0 does all the work. (This example, and most of the discussion in the rest of this section, leans heavily on chapter 2 of Russell 2008.)

But this won't do as a conception of analyticity either, because of the examples of the necessary a posteriori. Consider the example *Gold has atomic number 79*. It is true in virtue of what it means, that gold has atomic number 79, and how the world is. But it doesn't matter which world we choose; in any world gold has atomic number 79. Yet it is not, intuitively, analytic.

Russell suggests a solution to this problem that draws on the developments in externalist theories of meaning that we discussed in the introduction. Start with the following three-way distinction. (These definitions are a quote from page x of Russell 2008.)

- *Character*: The thing speakers must know (perhaps tacitly) to count as understanding an expression.
- *Content*: What the word contributes to what a sentence containing it says (the proposition it expresses).
- *Reference Determiner*: A condition which an object must meet in order to be the reference of, or fall in the extension of, an expression.

These can all come apart. In the case of pure indexicals like *I*, the content comes apart from the character and reference determiner in familiar ways. But it is tempting in those cases to equate character and reference determiner. What makes it the case that a token of *I* picks out me is that I use it, and that relation between usage and content is what someone must know to understand the term. But that's an all-too-special case. I can be a competent user of the name "Alex" as a name for my friend Alex without knowing whether she got that name at birth in the normal way, or knowing whether she acquired it later. Competence may require that I know the reference of Alex was somehow determined to be her, but I need not know what that reference determiner was.

Moreover, the character and reference determiner relate to contexts in different ways. It is a familiar point that a sentence like *If you were speaking, I would have been speaking* may be false. That's because when we evaluate the *I* in the consequent, we don't look to who the speaker is in the context of evaluation, that is, the world where you are speaking, but to the world of utterance, that is, the context of my utterance. Just like this familiar distinction between contexts of utterance and contexts of evaluation, Russell requires us to think about contexts of introduction. An example helps bring this out.

Say I, on Monday, introduce the name "Inigo" for the shortest sword fighter. When the name is used on Tuesday, it need not pick out the shortest sword fighter, even in the context of utterance. Inigo might have grown, or a shorter person may have taken up sword fighting. Of course, when I use the name in counterfactuals, it might pick out someone who was never a sword fighter. So we need to distinguish the context the term was introduced in, in this case Monday, from the context it is uttered in, in this case Tuesday.

With these distinctions in mind, we can give Russell's first pass at a definition of analyticity:

> A sentence *S* is true in virtue of meaning just in case for all pairs of context of introduction and context of utterance, the proposition expressed by *S* with respect to those contexts is true in the context of evaluation. (Russell 2008, 56)

This will, says Russell, solve the problem about gold. It is true that when someone now utters *Gold has atomic number 79*, they express a necessary truth. But we could have introduced the terms in the very same way, and had the world not cooperated, this sentence would have been false. Indeed, Russell splits analyticity from necessity twice over. She thinks that *I am here now* will be analytic in this sense though it expresses a contingent proposition.

This does rely on understanding what it is for terms in different worlds to have the same reference determiner. Perhaps one could object that had we been pointing at something else when we introduced the term *gold*, we would have been using a crucially different reference determiner. But the issues here about the metaphysics of words and demonstrations, are subtle, and Russell's view that the same reference determiner could determine different contents in different worlds seems plausible.

Russell says that this is a perfectly good notion of truth in virtue of meaning. Of course, as she says, it is really a kind of truth in virtue of what determines meaning, not meaning itself. Reference determiners are part of meta-semantics, not semantics. But that seems continuous enough with the tradition. And there is no reason to think that analytic sentences, so understood, will be epistemologically distinctive. In this respect we may end up agreeing with the primary conclusion of the discussion of metaphysical analyticity in chapter 3 of Williamson (2007), namely that it isn't directly relevant to philosophical methodology. But it could be an interesting notion in its own right, and as discussed in section 13.7, it could be philosophically useful without playing its traditional role.

I have considerably simplified the presentation of Russell's view, however. The definition so far implies that some theorems of geometry, and perhaps fundamental laws of ethics, will be analytic. Like Kant, Russell wants these to be synthetic. Her solution is to say that analytic truths must not just retain their truth value as we change the context of introduction, utterance and evaluation, but that they must do so because the reference determiners of their parts stand in the right kind of containment relations. But spelling this part of her view out will take too much space, so instead I'll close with a discussion of epistemological analyticity.

9. Epistemological Accounts of Analyticity

In a series of influential articles, Paul Boghossian (1996; 1997; 2003) argued that we should accept Quine's argument against metaphysical versions of the analytic/synthetic distinction, but that Quine's arguments left untouched an epistemic understanding of the distinction. On this way of understanding the distinction, a sentence is analytic iff it is knowably true merely in virtue of understanding it. Consider, for instance, this sentence.

(E) If frogs bark and ducks howl, then frogs bark.

Now make the following four assumptions:

1. Understanding the non-logical terms in (E) suffices to see it is of the form $(A \wedge B) \rightarrow A$.
2. For logical terms like $\wedge$ and $\overset{\cdot}{\rightarrow}$, understanding involves accepting, perhaps implicitly in one's inferential practices, the basic introduction and elimination rules they license.
3. For $\wedge$, the basic rules are the familiar introduction and elimination rules.
4. For $\rightarrow$, the basic rules are modus ponens and conditional proof.

Then anyone who understands (E) is in a position to prove it to be true by a trivial three-line proof. Generalizing this example, we can get an argument that all logically true sentences are analytic. Generalizing a bit further, we may be able to argue that the propositions they express are a priori knowable, but this requires resolving many of the issues we have already discussed in the discussion of the a priori, and I will set it aside for the remainder of this chapter.[9]

The problem we will focus on is that assumptions 2 and 4, and hence presumably 3, are not clearly true. Actually 4 as stated is almost surely false. If the basic rules for → are modus ponens and conditional proof, then → is material implication. But → was meant to be our symbol for natural language "if", which is not material implication. So the rule must be something else. It isn't clear what this rule could be. It is plausible that we can use a restricted version of conditional proof when reasoning about "if", such as a version which requires that there be no undischarged assumptions when we apply conditional proof. That will make the proof of (E) go through, but it is unlikely to be a basic rule in the relevant sense, since it does not combine with an elimination rule (i.e. modus ponens) to pick out a unique meaning for "if".

Disagreement about the introduction rule for "if" is endemic to the literature on conditionals. But there is almost a consensus that the elimination rule is modus ponens. Almost, but not quite—Vann McGee (1985) is a notable dissenter. Timothy Williamson (2007 ch. 4) uses the existence of notable dissenters like McGee to mount a sustained assault on Boghossian's position. It is a consequence of the assumptions we have made, and which Boghossian needs, that anyone who doesn't accept modus ponens does not understand "if". But that seems implausible. By any familiar standard, McGee understands conditionals quite well. Indeed, he is an expert on them.

This point generalizes, as Williamson stresses. On the inferentialist view about the meaning of logical terms, in any debate about the correctness of fundamental logical principles, either one party doesn't understand the key terms, or the parties are speaking at cross purposes. The intuitionist mathematician endorses the sentence, "All functions are continuous", and the classical mathematician rejects it. But it isn't plausible that one party fails to understand "all", "functions", or "continuous", or that they are speaking at cross purposes in that they are assigning different meanings to one of these terms. (I'm assuming the context makes it clear that both parties are speaking of functions whose domain is the reals, and whose range is a subset of the reals.) I've used an example from real analysis, but we could make the same point less pithily using Peirce's Law if we wanted to stick to propositional logic.

I'll close with two replies on behalf of the defending of epistemic analyticity, and some reasons for being dissatisfied with each. The discussion will follow somewhat the recent exchange between Boghossian (2011) and Williamson (2011).

The first response says that we shouldn't have said if a sentence is epistemically analytic, then understanding it is sufficient for knowing that a sentence is true. Rather, we should have said that knowing the meaning is sufficient for knowing the sentence is true. Notably,

[9] There is an interesting worry around here that the three-line argument for (E) is circular, and so cannot justify (E), and this fact undermines Boghossian's argument that it is a priori. See Ebert (2005) and Jenkins (2008) for two ways of developing this worry.

Boghossian does not make this defence in his exchange with Williamson, so it seems he accepts that Williamson was right to take epistemic analyticity to involve a connection between understanding and knowability. And this seems to be right. Consider again *Water contains oxygen*. In one sense of meaning, the meaning of "water" is H_2O. So anyone who knows what "water" means in that sense knows that *Water contains oxygen* is true. But it doesn't feel like this claim is analytic, especially not in the epistemic sense that interests Boghossian. It is possible that there is some other sense of meaning that will be more useful for Boghossian's project, but it isn't clear that knowing the meaning in this other sense will differ particularly from understanding.

The second response, and one that Boghossian has used on several occasions, is that the only plausible theory of meaning for the logical connectives is inferentialist, and on an inferentialist theory of meaning it will be true that anyone who understands a connective is disposed to reason correctly with it. That last sentence is deliberately sloppy, much more so than any statement of the response in Boghossian's own work. But the sloppiness is there because it makes a potential equivocation more easily visible.

Consider a theory that says the meaning of a logical connective is either constituted by, or at least constitutively connected to, its appropriate inferential rules. But to understand the term is not to grasp the meaning in this sense, any more than to understand the term "water" one has to know it is H_2O. Rather, understanding involves participating in the right kind of way in a social practice, and it is that social practice (plus perhaps some facts about the nature of logic, if such facts there be) that determines the appropriate inferential rules for the connective.

Is the theory in the previous paragraph inferentialist? If not, then it is false that no theory other than inferentialism is plausible as an account of the meaning of the logical connectives. For this kind of socialized theory of meaning is, it seems to me, highly plausible. (Williamson (2011) notes that a socialized theory of meaning for the connectives is plausible, though I don't think he would sign up for the view that the result of such socialization is a theory in terms of inferential rules.) If, on the other hand, the theory is inferentialist, then it doesn't follow that understanding requires a disposition to use the rules. Perhaps understanding requires being part of a community many members of which have the appropriate dispositions, but it does not require that any one member have these dispositions. So it won't be true that mere understanding puts one in a position to know. At best, understanding a logical truth means one is in a community in which some people are in a position to the sentence is true. But that doesn't do much to rescue the notion of epistemic analyticity.

References

Ayer, Alfred. 1936. *Language, Truth and Logic*. London: Gollantz.

Block, Ned, and Robert Stalnaker. 1999. "Conceptual Analysis, Dualism, and the Explanatory Gap." *Philosophical Review* 108 (1): 1–46.

Boghossian, Paul A. 1996. "Analyticity Reconsidered." *Noûs* 30 (3): 360–91.

Boghossian, Paul A. 2003. "Epistemic Analyticity: A Defense." *Grazer Philosophische Studien* 66 (1): 15–35.

Boghossian, Paul A. 2011. "Williamson on the a Priori and the Analytic." *Philosophy and Phenomenological Research* 82 (2): 488–97.

Boghossian, Paul Artin. 1997. "Analyticity." In *A Companion to the Philosophy of Language*, edited by Bob Hale and Crispin Wright, 331–68. Oxford: Blackwell.

Cappelen, Herman. 2012. *Philosophy Without Intuitions*. Oxford: Oxford University Press.

Chalmers, David. 1996. *The Conscious Mind*. Oxford: Oxford University Press.

Chien, Sarina Hui-Lin. 2011. "No More Top-Heavy Bias: Infants and Adults Prefer Upright Faces but Not Top-Heavy Geometric or Face-Like Patterns." *Journal of Vision* 11 (6): 13, 1–14.

Cohen, Stewart. 2010. "Bootstrapping, Defeasible Reasoning and *a Priori* Justification." *Philosophical Perspectives* 24 (1): 141–59.

Conee, Earl, and Richard Feldman. 2004. *Evidentialism: Essays in Epistemology*. Oxford: Oxford University Press.

Davies, Martin, and I. L. Humberstone. 1980. "Two Notions of Necessity." *Philosophical Studies* 38: 1–31.

Dogramaci, Sinan. 2010. "Knowledge of Validity." *Noûs* 44: 403–32.

Ebert, Philip. 2005. "Transmission of Warrant Failure and the Notion of Epistemic Analyticity." *Australasian Journal of Philosophy* 83: 505–22.

Evans, Gareth. 1979. "Reference and Contingency." *The Monist* 62: 161–89.

Hawthorne, John. 2002. "Deeply Contingent a Priori Knowledge." *Philosophy and Phenomenological Research* 65: 247–69.

Hawthorne, John. 2007. "Craziness and Metasemantics." *Philosophical Review* 116 (3): 427–440.

Heron-Delaney, Michelle, Sylvia Wirth, and Olivier Pascalis. 2011. "Infants' Knowledge of Their Own Species." *Philosophical Transactions of the Royal Society B* 366 (1571): 1753–63.

Hume, David. 1978. *A Treatise on Human Nature*. Edited by L. A. Selby-Bigge and P. H. Nidditch. Second. Oxford: Clarendon Press.

Ichikawa, Jonathan, and Benjamin Jarvis. 2009. "Thought-Experiment Intuitions and Truth in Fiction." *Philosophical Studies* 142 (2): 221–46.

Jackson, Frank. 1998. *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Clarendon Press: Oxford.

Jehle, David, and Brian Weatherson. 2012. "Dogmatism, Probability and Logical Uncertainty." In *New Waves in Philosophical Logic*, edited by Greg Restall and Gillian Russell, 95–111. Basingstoke: Palgrave Macmillan.

Jenkins, C. S. 2008. "Boghossian and Epistemic Analyticity." *Croatian Journal of Philosophy* 8 (22): 113–27.

Kant, Immanuel. 1999. *Critique of Pure Reason*. Edited by Paul Guyer and Allen Wood. Cambridge: Cambridge University Press.

Kaplan, David. 1989. "Demonstratives." In *Themes from Kaplan*, edited by Joseph Almog, John Perry, and Howard Wettstein, 481–563. Oxford: Oxford University Press.

Kitcher, Philip. 1980. "A Priori Knowledge." *Journal of Philosophy* 89: 3–23.

Kripke, Saul. 1980. *Naming and Necessity*. Cambridge: Harvard University Press.

McGee, Vann. 1985. "A Counterexample to Modus Ponens." *Journal of Philosophy* 82: 462–71.

Pargetter, Robert, and John Bigelow. 1997. "The Validation of Induction." *Australasian Journal of Philosophy* 75 (1): 62–76.

Putnam, Hilary. 1973. "Meaning and Reference." *Journal of Philosophy* 70: 699–711.

Putnam, Hillary. 1981. *Reason, Truth and History*. Cambridge: Cambridge University Press.

Quine, W. V. O. 1936. "Truth by Convention." In *Philosophical Essays for a. N. Whitehead*, edited by O. H. Lee, 90–124. New York: Longmans.

Quine, W. V. O. 1951. "Two Dogmas of Empiricism." *Philosophical Review* 60 (1): 20–43.

Quine, W. V. O. 1960. *Word and Object*. Cambridge, MA: MIT Press.
Russell, Gillian. 2008. *Truth in Virtue of Meaning: A Defence of the Analytic/Synthetic Distinction*. Oxford: Oxford University Press.
Sober, Elliot. 2000. "Quine's Two Dogmas." *Aristotelian Society Supplementary Volume* 74 (1): 237–80. doi:10.1111/1467-8349.00071.
Strang, Nicholas F. 2011. "Did Kant Conflate the Necessary and the *a Priori*." *Noûs* 45 (3): 443–71. doi:10.1111/j.1468-0068.2010.00809.x.
Thomson, Judith Jarvis. 1971. "A Defense of Abortion." *Philosophy and Public Affairs* 1 (1): 47–66.
Weatherson, Brian. 2005. "Scepticism, Rationalism and Externalism." *Oxford Studies in Epistemology* 1: 311–31.
Weatherson, Brian. 2007. "The Bayesian and the Dogmatist." *Proceedings of the Aristotelian Society* 107: 169–85.
Weatherson, Brian. 2012. "Induction and Supposition." *The Reasoner* 6 (6): 78–80.
White, Roger. 2006. "Problems for Dogmatism." *Philosophical Studies* 131: 525–57.
Whitmore, Charles E. 1945. "Mill and Mathematics: An Historical Note." *Journal of the History of Ideas* 6 (1): 109–12.
Williamson, Timothy. 2000. *Knowledge and its Limits*. Oxford: Oxford University Press.
Williamson, Timothy. 2007. *The Philosophy of Philosophy*. Oxford: Blackwell Pub. Ltd.
Williamson, Timothy. 2011. "Reply to Boghossian." *Philosophy and Phenomenological Research* 82 (2): 498–506.
Williamson, Timothy. 2013. "How Deep Is the Distinction Between a Priori and a Posteriori Knowledge." In *The a Priori in Philosophy*, edited by Albert Casullo and Joshua C. Thurow, 291–312. Oxford: Oxford University Press.
Yablo, Stephen. 2002. "Coulda, Woulda, Shoulda." In *Conceivability and Possibility*, edited by Tamar Szabó Gendler and John Hawthorne, 441–92. Oxford: Oxford University Press.
Zagzebski, Linda. 1994. "The Inescapability of Gettier Problems." *The Philosophical Quarterly* 44 (174): 65–73.

CHAPTER 14

PHILOSOPHICAL AND CONCEPTUAL ANALYSIS

JEFFREY C. KING

1. Introduction

Philosophers spend a lot of time attempting to give analyses of philosophically interesting notions. Analyses have been proposed for knowledge, moral rightness, species-hood, object-hood, persistence, change, reference, and much more. It is therefore surprising that there isn't more consensus among philosophers regarding what they are attempting to do in providing purported analyses. Philosophers don't agree about the things that are being analyzed, nor what it is to analyze something. In what follows, we'll see a sampling of views on what philosophical analysis is. The present work isn't meant to be exhaustive and there is much work that will not be discussed. However, it does purport to illustrate the main lines of thinking about analysis in recent philosophy. The present work also makes no attempt to discuss the views about analysis of historical figures like Gottlob Frege, G. E. Moore, and Bertrand Russell. There is a rich literature on this topic and interested readers should consult it. Here, we focus on more contemporary views.

2. Views of Philosophical Analysis that Address the Paradox of Analysis

In this section we discuss views of philosophical analysis that aim to address the so-called *paradox of analysis*. Though the paradox is often said to have been formulated by Plato, contemporary interest in it is due to the formulations of the paradox by C. H. Langford and G. E. Moore. Hence, we do well to start here. What follows is Moore's influential statement of the paradox:

> But, now, if we say, as I propose to, that to make any of the above three statements[1] is to give an analysis of the concept 'brother', we are obviously faced with a puzzle which Mr. Langford calls 'the paradox of analysis'. Suppose we use still another way, a fourth way, of expressing the very same statement which is expressed in those three ways I gave, and say: 'To be a brother is the same thing as to be a male sibling.' The paradox arises from the fact that, *if* this statement is true, then it seems that you would be making exactly the same statement if you said: 'To be a brother is the same thing as to be a brother.' But it is obvious that these two statements are *not* the same; and obvious also that nobody would say that by asserting 'To be a brother is to be a brother.' you were giving an analysis of the concept 'brother'.[2]

Moore claims that if the following is a correct analysis:

1. To be a brother is to be a male sibling

then if you say either of the two following sentences, you would be making the same statement and, in both cases, giving an analysis of the concept 'brother':

2a. To be a brother is to be a brother
2b. To be a brother is to be a male sibling

But, Moore says, both of these things are obviously false: saying 2a and 2b does not amount to making the same statement; and saying 2a does not amount to giving an analysis.

In effect, Moore is claiming that if 1 expresses an analysis, then the sentences 2a and 2b stand in some relation (saying 2a is making the same statement as saying 2b; in saying 2a or 2b, one is giving an analysis). But there is strong reason to think 2a and 2b do not stand in this relation. Hence, in its most general form, the paradox can be viewed as follows. We begin by supposing we have a correct analysis:

1. To be a brother is to be a male sibling

Next, it is alleged that if 1 is a correct analysis, then the following sentences stand in some relation:

2a. To be a brother is to be a brother
2b. To be a brother is to be a male sibling

What relation they are claimed to stand in varies with the version of the paradox: it may be claimed that they express the same proposition; that they mean the same thing; that they are both analyses and so on. Reasons are then given for the claim that they cannot stand in this relation. Thus the paradox: assuming that 1 is an analysis, 2a and 2b both must and must not stand in some relation.

Of course, any view of philosophical or conceptual analysis that addresses the paradox of analysis must somehow resolve it. But presumably theories addressing the paradox should

[1] The statements in question are: 'The concept "being a brother" is identical to the concept "being a male sibling" '; 'The propositional function "x is a brother" is identical to the propositional function "x is a male sibling".'; and 'To say that a person is a brother is the same thing as to say that that person is a male sibling.' See Schilpp 1942 p. 665.

[2] Schilpp 1942 p. 665.

do more than this. For example, the analysis given in 1 is importantly different from both of the following two claims, which appear to be (purported) analyses in some sense as well:

> (JTB) To be in instance of knowledge is to be an instance of justified true belief(H_2O)
> To be water is to be H_2O

In particular, 1 is uninteresting and trivial in a way that JTB and H_2O are not. Further, JTB and 1 can be the result of armchair theorizing. Not so for H_2O. An account of analysis that addresses the paradox of analysis should explain these things. Let's now turn to such accounts.

We begin with the view of Sosa 1983. Sosa is very much concerned with the paradox of analysis, and states it as follows:

> (A) To be a cube is (=) to be a closed solid with sides all square.
>
> (C1) No one can think consciously of being a cube without thinking consciously of being a closed solid with all sides square.
>
> (C1) follows from A given:
>
> RCT: Thinking (in various modes) is a relation between a thinker and a Thought (in an extension of Frege's sense which covers not only propositions but also properties, such as being a cube).[3]

Sosa supposes A to follow from the fact that the analysis of what it is to be a cube is that it is to be a closed solid with all sides square. RCT looks quite plausible. But C1 looks false. According to Sosa 1983, it is complex properties, built up out of other properties and relations, that are the objects of analysis.[4] An analysis specifies the constituents of the property being analyzed and how they are combined to yield the property being analyzed. Sosa refers to the latter as a 'logical mode of constitution', and explicitly mentions negation and conjunction.[5] Sosa also considers identity to be a limiting case of a mode of constitution, so that every property is constituted out of itself by the mode of constitution of identity. Hence, in general, an analysis is expressed by a sentence of the form 'P is constituted by mode of constitution M out of constituents $C_1, \ldots, C_n$ in that order.' An example Sosa uses is this: the property of being a cube is constituted by the mode of constitution *conjunction* out of constituents *being a closed solid* and *having all sides square*.

But how does all this solve the paradox of analysis that Sosa originally stated? Well, it doesn't really. Sosa thinks the property of being a cube and the property of being a closed solid with all sides square just are the same property. And given his commitments, this seems right: after

[3] Sosa 1983 pp. 695–696. Sosa also formulates a version of the paradox that begins with an identity claim about propositions instead of the identity claim about properties (A) above. I'll stick to A and C1 since Sosa spends more time on this version. Thinking back to my general schema for the paradox of analysis, we can say that on Sosa's version, assuming that *being a closed solid with all sides square* provides an analysis of the property of *being a cube*, then the property identity A is true (the properties must stand in some relation—identity; of course on my version above, it was *propositions* or *sentences* that must stand in some relation assuming some analysis is correct). But then that the apparently false C1 follows from A (given RCT) gives us reason to think that the properties in question are not identical contra A.

[4] Actually, what are analyzed on Sosa's view are Thoughts, which include properties and propositions. But as I indicated I'll concentrate on properties here.

[5] See p. 705.

all, if the property of being a cube is composed by conjunction out of the properties of being a closed solid and having all sides square, it is hard to see how the property of being a cube could fail to be the same property as the property of being a closed solid with all sides square. Sosa also thinks that the English expressions 'cube' and 'closed solid with all sides square' express the same property. Hence, C1 above is *true* and *does follow* from A and RCT. Our reluctance to accept C1, Sosa claims, stems from confusing it with the following false claim:

> C1' No one can consciously think of *being a cube* as bearing the relation of identity to *being a cube* (or *being a cube* being constituted out of *being a cube* by *identity*) without thinking of *being a cube* as the bearer of *conjunction* to the properties of being a closed solid and having all sides square.

Sosa's account has many merits. The idea that analyses state how the analyzed thing is composed out of other things and what the mode of composition is is powerful and has intuitive appeal. Further, that Sosa's account entails that being a cube and being a closed solid with all sides square just are the same property, and more generally that on his view in giving analyses one is trying to say what the thing being analyzed *is* are virtues of the view. One difficulty with Sosa's view is that he doesn't explain why it is that we confuse the true C1 with the false C1'. Further, being forced to an error theory on which we are always wrong about the truth-value of sentences like C1 is a significant cost. Finally, Sosa doesn't say what distinguishes *philosophical* analyses from other things that appear to be analyses in his sense. A1 after all is not a philosophical analysis. Further, recall the example of H_2O we have discussed. It is plausible that this tells us the constituents of the property of being water and how they are combined in the complex property of being water. So this appears to be an analysis in Sosa's sense. But it clearly is not a philosophical analysis. One would want an account of what makes an analysis philosophical. Similar remarks apply to 'To be a brother is to be a male sibling', which again appears to be an analysis in some sense, but not a philosophical analysis.

Ackerman (1981, 1986) gives an account of analysis on which properties are the objects of analysis.[6] Ackerman's (1986) version of the paradox of analysis is that if the following is a correct analysis:

> (3) To be an instance of knowledge is to be an instance of justified true belief

then the property of being knowledge must just be the property of being a justified true belief. But then 3 expresses the same proposition as

> (3a) To be knowledge is to be knowledge

But of course there is reason to think 3 and 3a do not express the same proposition (3a is trivial; 3 can be informative, etc.). Ackerman blocks the paradox by formulating an account of analysis according to which when a property P provides the analysis of a property Q, P and Q are nonetheless distinct properties.[7] Hence, 3 and 3a do not express the same proposition on Ackerman's view. Of course, the pressing question for such a view is: what relation

[6] Ackerman 1981 talks of concepts being the objects of analysis. Ackerman 1986 talks instead of properties and makes clear that she uses 'concept' and 'property' interchangeably (p. 306).

[7] Chisholm and Potter 1981 is another account on which when P analyzes Q, P and Q are distinct (Chisholm and Potter take properties to be the objects of analysis). Chisholm and Potter define a

does property P bear to the distinct property Q when P analyzes Q? The first two conditions are: (i) P and Q are necessarily coextensive; (ii) it is knowable a priori that P and Q are coextensive. However, as Ackerman acknowledges, these two conditions are not sufficient for something being a philosophical analysis since they don't rule out the property of being the fourth root of 1296 as providing an analysis of the property of being 6. The final condition for P to provide a philosophical analysis of Q is that the claim that necessarily to be Q is to be P can be justified by the 'philosophical example/counterexample method': we formulate hypothetical test cases and ask our subject K: 'Is this a case of Q?'. We then contrast the descriptions of cases in which K answers affirmatively with those in which she does not. *Generalizing* from these descriptions we arrive at the properties and their mode of combination that constitute the analysis of (K's notion of) Q.

One difficulty with this account is raised by Ackerman (1986) herself. Ackerman's third condition on analyses is designed to get beyond something that is merely a priori knowable to be coextensive with Q and necessarily coextensive with Q (on pain of things like *triangularity* being an analysis of *trilaterality*). However, suppose upon hearing K's answers as to which cases are Q and which aren't I *generalize* to P being the analysis of Q. Now take any property R≠P that is a priori knowable to be coextensive with Q and that is necessarily coextensive with Q. Given K's responses I could just as easily have *generalized* to R being the analysis of Q. After all, nothing has been said about how to generalize to P from K's answers as to which hypothetical cases are Q except that P must capture those answers. But if P does, then R will, since P, Q and R are necessarily coextensive (and a priori knowable to be coextensive). Hence, for all that has been said, Ackerman's third condition that was supposed to get beyond P being merely necessarily coextensive with Q and a priori knowable to be coextensive with Q fails to do so since it fails to distinguish between P (which, we can suppose, *does* analyze Q) and an arbitrary property R that is merely necessarily coextensive with Q and knowable a priori to be so.

A second difficulty is that the third condition essentially says that the claim that necessarily something is Q iff it is P can be justified by a certain kind of philosophical inquiry. Suppose the claim that necessarily something is Q iff it is P *can* be so justified (and satisfies Ackerman's other two conditions). Then the fact that it can be so justified is either a brute fact or supervenes on some other relation between P and Q. The former seems implausible: how could it be a brute fact that the claim that necessarily something is Q iff it is P be the possible outcome of a certain sort of philosophical inquiry? Hence it must be that there is some as yet undiscovered relation R that obtains between P and Q and that explains *why* the claim that necessarily something is Q iff it is P can be justified in the way described. In that case, arguably that P and Q stand in R ought to be Ackerman's third condition on philosophical analyses. But then Ackerman hasn't really given us an account of analysis. Instead she has told us that P analyzes Q iff (i) P is necessarily coextensive with Q; (ii) it is knowable a priori that P is coextensive with Q; and (iii) there is some relation R between P and Q that explains why the claim that necessarily something is Q iff it is P can be justified

number of quasi-logical notions and use them to define how P and Q have to be related for P to analyze Q. However, in giving their definitions of the notions that are used to define what an analysis is, Chisholm and Potter rely on a number of undefined notions (specifically, conceiving, and attributing) whose application I simply don't understand. Thus I am unable to determine whether pairs of properties stand in the relation defined using these notions or not.

by the philosophical method described. Until we are told what R is, we really don't have an account of philosophical analysis. One way to see this is that Ackerman's third condition was designed to distinguish philosophical analyses from things like

(4) For all x, x is 6 iff x is x is the fourth root of 1296

that satisfy Ackerman's other two conditions on philosophical analyses. It is very unsatisfying to be told that 3 is a philosophical analysis and 4 isn't because though they both satisfy conditions 1 and 2 on philosophical analyses, only 3 can be the outcome of a certain type of philosophical inquiry. Of course, what is needed here is an explanation of *why* this is so.

Related to this, as with Sosa's account, nothing in Ackerman's account tells us how philosophical analyses differ from things like 'To be a brother is to be a male sibling' and 'Water is H_2O', which, as indicated, also appear to be analyses in some sense.

A final difficulty with Ackerman's view is shared by all views on which when P analyzes Q, P and Q are distinct properties. This just seems not to capture what we take ourselves to be doing in attempting to analyze some property. In so doing, we take ourselves to be trying to say what the property *is*, as I mentioned in discussing Sosa 1983. We do not take ourselves to be trying to come up with a property *intimately related* to the property we are trying to analyze. Perhaps if there were no plausible account of analysis on which in some sense the analyzing property and analyzed property are the same, we would be forced to an account on which these properties are different. But it doesn't appear to be a view we should embrace otherwise.

Jeffrey C. King's [1998, 2007] account of analyses begins by assuming that analyses will be stated by means of universally quantified biconditionals:

∀x (x is knowledge iff x is a justified true belief).

For simplicity, King assumes that the term of the left of the biconditional is syntactically simple and the term on the right is syntactically complex. King's framework for resolving the paradox of analysis comprises three elements, which King claims can be independently motivated. That this is so, King claims, makes his resolution of the paradox attractive.

The first element is the claim, endorsed by Sosa as we have seen as well, that some properties and relations are complex and are "made up" of other properties and relations. To take a very simple example, the property of being a brother is complex and is made up out of the properties of being male and being a sibling. These properties are combined conjunctively in the bachelor property.[8] Call the properties and relations that combine to make up a complex property or relation the latter's *components*.

The second element of King's framework is some version of a view about propositions defended in King [2007, 2009]. What needs to be assumed here is that the structure of a proposition is very close to the syntactic structure of the sentence expressing it at the level of syntax where quantifier scope is explicitly represented. In effect, this means that a proposition is a structured entity whose structure is very much like the structure of the sentence expressing it, where the semantic values of the words in the sentence are structured in the

[8] These latter properties may themselves be complex.

proposition more or less the way the words with those semantic values are structured in the sentence.

The final element of King's framework is the claim that there are at least three categories of words, where what distinguishes the words in each category from words in the others are the standards of competence governing the words. In category one are words that express complex properties and relations, where to be competent with the words requires one to know the component properties and relations and how they are combined to yield the complex property or relation expressed by the word. 'Bachelor' is a paradigm of a category one word: to be competent with it one must know that 'bachelor' expresses a complex property that results from conjunctively combining *being male* and *being unmarried*.[9] Category two words are words that express complex properties or relations where competence with the word requires one to be able to say whether the property is instantiated or not in a (possibly hypothetical) situation, given sufficient information about the situation and given that the situation is one in which the property is *paradigmatically* instantiated or not.[10] Paradigms of category two words are things like 'knowledge', 'reference', 'chair' etc. Finally, competence with words in category three neither requires speakers to be able to articulate the component properties and relations of the complex property or relation expressed by the word, nor properly apply the word in clear (possibly hypothetical) cases. It is hard to say exactly what competence *does* require here. Category three words include words that many have called *natural kind words* such as 'aluminum', 'elm', and so on.

Now on King's view of analysis, it is complex properties and relations that are the objects of analysis. In stating an analysis, one is saying what the component properties and relations are that make up the complex property being analyzed and how they are combined to form this complex property. Given King's assumption about the syntactic structure of sentences stating analyses, and his assumptions about the structures of propositions expressed by sentences, a sentence expressing a purported analysis and the proposition it expresses will look as follows:

S_a [[Every x][[R(x)] iff [C(x)]]]

P_a [[Every* x[[R*(x)] iff [$\mathscr{C}$(x)]]]

where Every* is the semantic value of 'Every', R* is the semantic value of 'R' (and is the property being analyzed) and $\mathscr{C}$(x) is what the complex predicate 'C(x)' contributes to the proposition expressed by S_a. Now according to King, in the case of entities like $\mathscr{C}$(x) that are contributed to propositions by syntactically complex predicates, in the definition of truth for propositions such entities will be mapped to properties. Take a simple case like the following proposition:

[[Mary*[hit* [Bill*]]]

[9] Of course ordinary competent speakers wouldn't put things this way, but that is how we as theorists would describe them, given our commitment to the first element of King's framework.

[10] The requirement that the situation be one in which the property in question is paradigmatically instantiated or not reflects the idea that competence with the word in question requires the speakers to "get it right" in cases in which the word clearly applies or clearly doesn't. Also, since category one words likely satisfy the condition stated for category two, we should add that no category one words are in category two.

where [[hit*[Bill*]] is the $\mathscr{C}$(x)-like complex entity contributed to this proposition by the complex predicate 'hit Bill' in the sentence 'Mary hit Bill'. In the definition of truth for propositions, this entity must be mapped to the property of *hitting Bill* and then the above proposition is true iff Mary possesses this property. In such a case, King says that [[hit*[Bill*]] *represents* the property of hitting Bill. Further, let's call hit* and Bill* the *constituents* of the complex, subpropositional $\mathscr{C}$(x)-like entity [[hit*[Bill*]]. Finally, note that it is plausible that hit* and Bill* are *components* of the presumably complex (and relational) property of *hitting Bill*. Exactly similar remarks apply to $\mathscr{C}$(x) in P_a: it will be mapped to a property by the definition of truth for propositions and it will have constituents. We can now state King's account of analysis: a proposition of the form P_a is an (correct) analysis iff (i) the property R* is identical to the property represented by $\mathscr{C}$(x); and (ii) the constituents of $\mathscr{C}$(x) are components of the property R*.

This does not yet distinguish *philosophical* analyses from (what are arguably on King's account) analyses like the uninformative analysis 1 and the 'scientific analysis' H_2O above. King claims that philosophical analyses are distinguished from the former in terms of the sort of epistemic relations typical members of a linguistic community bear to the property being analyzed. In turn, King thinks that the epistemic relations typical members of a linguistic community bear to a property are reflected in the standards of competence associated with the word expressing the property in the community. Hence, a proposition P is a philosophical analysis for a linguistic community C iff (i) P is an analysis; and (ii) there is a sentence of the language of C that expresses P, and the word that contributes the analyzed property to P belongs to category two.[11] King's idea is that if the word expressing the analyzed property belongs to category one or three, the result will be an uninformative analysis or a scientific analysis, respectively. Further, King thinks that if one is a member of a linguistic community C where the word w expressing a given property in the language of C belongs to category two, speakers' competence with w in C allows speakers to consider hypothetical situations and determine whether the property expressed by w is instantiated in those situations or not. This puts speakers in a position to formulate hypotheses about the analysis of the property expressed by w.

As to problems with King's view, one might think that true statements such as 'To be good is to be pleasurable.'[12] are philosophical analysis. But on King's view, they won't be since the predicate doesn't have constituents whose semantic values are components of the property being analyzed. Hence, there is no sense in which the property of being good is analyzed here on King's view of analysis. A second worry concerns how we could ever know an analysis is correct on King's account. Given a purported analysis of the form of S_a above, suppose we could come to known that it is necessarily true. Still, how could we ever determine that the further conditions required for a successful analysis are met? Really, this amounts to asking how we could know that we have successfully identified the components (and how they are combined) of a complex property or relation. This epistemological worry applies to any account on which analyses purport to tell us the components of a complex property or relation and how they are combined in it.

[11] For simplicity King assumes here that there are not two words or more words of the language of C belonging to different categories that express the analyzed property.

[12] Suppose, counterfactually, that this is a true statement.

3. Views of Analysis That Are Not Aimed at Addressing the Paradox of Analysis

The views we have looked at so far have been views of philosophical analysis that try to explain the paradox of analysis and the felt difference between 3 and 3a as set out earlier. However, there are views of analysis that are aimed at addressing quite different issues.

One such view is that of David Lewis (1966, 1970, 1994). Lewis (1994) holds that it is an a priori truth that there are fundamental (perfectly natural) properties and relations, and that every contingent truth must be made true by the pattern of instantiation of these fundamental properties and relations. As he puts it, 'The whole truth about the world, including the mental part, supervenes on this pattern.'[13] Lewis further takes the fundamental properties and relations to be physical. This claim Lewis calls *materialism*. Of course there is no a priori guarantee that materialism is correct. But putting together the a priori claim that everything supervenes on the pattern of instantiation of the fundamental properties and relations, with the a posteriori claim that the fundamental properties and relations are all physical, we get the a posteriori claim that everything supervenes on the pattern of instantiation of physical properties and relations. In particular, the mental items so supervene.

But this raises a puzzle for Lewis. Some features of the world that supervene on the physical properties and relations will be given by extremely complex physical conditions that are too complex for beings like us to entertain or track. But mental features aren't like that. We seem to do a surprisingly good job of tracking the beliefs, desires, and so forth of others. This must mean, Lewis thinks, that there is a certain kind of simplicity about mental phenomena when looked at in the right way. It is the job of conceptual analysis to reveal that simplicity.

Here Lewis appeals to a largely tacit theory we all grasp of how we work mentally. Lewis call this theory *folk psychology*. Lewis thinks folk psychology is largely correct and that it is a powerful instrument for predicting and explaining human behavior. For any given mental state M, folk psychology will have principles telling us the causal relations between M, perceptual input, behavioral output and other mental states (e.g. it might tell us very roughly that pain is caused by certain perceptual inputs, causes certain behavioral outputs, and leads to being in other mental states like anger, etc). Imagine that we conjoin all the principles of folk psychology and let M_1, … M_n be names of all the mental states the theory mentions.[14] Let this sentence be 'T[M_1, … ,M_n]'. This, in effect, gives us our simultaneous conceptual analysis of the mental states in our folk psychology by assigning to the n-tuples of these states a joint causal role, including causal relations between the mental states

[13] See p. 292. Lewis' gloss on the supervenience of everything on the instantiation of perfectly natural properties and relations is on that same page: 'If two possible worlds were exactly isomorphic in their patterns of coinstantiation of fundamental properties and relations, they would thereby be exactly alike *simpliciter*.'

[14] For an explanation of why we can use *names* of mental states and why if the conjunction is infinitely long there is no problem, see Lewis 1970 p. 80.

(as well as causal relations to perceptual input and behavioral output). We can think of '$T[M_1, \ldots, M_n]$' as saying 'the states that typically occupy the $M_1 \ldots, M_n$ roles are typically causally related to each other, perceptual input and behavioral output as follows: …'. Now it is an open question and an empirical question what states do in fact occupy these roles. But if Lewis is right and every feature of the world supervenes on physical properties, the occupier of the roles are ultimately physical states. And this, Lewis thinks, gives us a simple argument for the view that mental states are ultimately identical to physical states. For simplicity, let's suppress the idea that the mental states $M_1, \ldots, M_n$ are interdefined and focus on a single mental state M. The argument that M is a physical state runs as follows:

Mental state M = the occupier of the M role (conceptual analysis)
The occupier of the M role = physical state P (empirical claim)
Therefore, mental state M = physical state P

It should be noted that the second premise here is contingent, according to Lewis. Hence, the conclusion is contingent as well. Some different physical state could have occupied the M role (or even some nonphysical state).[15] It is important to see that for Lewis, P *is* the state M. So had things been different, some other state would have been M.

In summary, for Lewis conceptual analysis is simply a means for picking out the physical state that occupies a certain role, where formulating what that role is constitutes a conceptual analysis of the relevant notion. This is done in the service of reconciling physicalism with mental features of the world. As a result, Lewis's view of conceptual analysis has no obvious application to the paradox of analysis.

Whatever the merits of Lewis' view here regarding mental states, the question arises whether Lewis' conception of conceptual analysis is broadly applicable to other philosophically interesting concepts. Most of the concerns center around what the theory is that is the analogue of folk psychology for the other philosophically interesting concepts one might hope to analyze. In the case of ethical concepts, for example, it is not clear that there is a consistent folk theory of morality, since there is wide-ranging moral disagreement. Further, in the case of virtually *any* philosophically interesting concept, one must be given some idea of how to determine what claims involving the concept, and related concepts, count as part of the relevant folk theory, as opposed to merely being claims involving the concept and related concepts.

A related view of conceptual analysis appears in the work of Frank Jackson (1994, 1998) and David Chalmers (1996). We'll put things in Chalmers' terms here, though Jackson's views in crucial respects are similar. It will be useful to put things in terms of the two-dimensional semantic framework Jackson 1998 and Chalmers 1996 employ.[16]

In this framework, expressions are associated with two functions from possible worlds to extensions. Such functions are generally called *intensions*. A word like 'water' has what is often called a *primary intension*, which maps a possible world to the stuff that is water in that world.[17]

[15] I suppress here Lewis' related idea that it may even be that for nonhuman animals or aliens (or even subpopulations of humans), something other than P does occupy the M role for them. In that case we would have to relativize to kinds of things K, and instead of the second premise above, we would have: M in kind K = P. See Lewis 1994 pp. 305–307.

[16] For criticism of the two-dimensional approach, see Soames 2004.

[17] Chalmers actually uses the set of *centered* possible worlds as the domain of primary intensions. For simplicity I'll ignore that here.

The idea here is that when the function is applied to a world, it gives as its output what the extension of 'water' would be if that world turns out to be actual. So consider a world where the oceans and lakes are filled with a chemical XYZ (which is not H_2O). XYZ also falls from the skies there during storms, comes out of faucets, and so on. The intuition here is that if the world turns out to be like that, water is XYZ. So the primary intension of 'water' maps the XYZ world to XYZ. Of course, it also maps the actual world to H_2O. To repeat, it maps any world to what would be water in that world if the actual world turns out that way.

But 'water' also has a secondary intension. This intension maps a possible world, now considered as counterfactual, to what is water at that world given that water is H_2O. When we consider the XYZ world as counterfactual, and ask what is water there, the answer is that water is H_2O there and that there is no water in the XYZ world. Given that water in the actual world is H_2O, water is H_2O in *every* possible world. Hence, the secondary intension of 'water' maps every possible world to H_2O. As we've just seen, the primary and secondary intentions are different for an expression like 'water'. This is because what we will say is water at a given world w depends on whether we are thinking of w as the way the actual world turned out (primary intension) or as a counterfactual world, holding the actual world, in which water is H_2O, fixed (secondary intension). However, for some expressions the primary and secondary intensions collapse. Consider 'square'. Whether we consider a world w as how the actual world turned out or as a counterfactual world (holding the actual world fixed) makes no difference to what we would say are squares at the world in question. So primary and secondary intentions collapse for such a word.

Now specifying the primary intension for a term Chalmers calls *conceptual analysis*. The primary intension encodes the way the secondary intension gets fixed given the way the actual world turns out to be. Hence, the primary intension encapsulates the application conditions for a term given a world considered as actual. In the case of 'water', very roughly speaking it applies to the local watery stuff in any world considered as actual. That is why it yields XYZ at the XYZ world and H_2O in the actual world. We discover what the primary intension of a term is by considering various ways the world might be and asking: 'If the world turns out that way, what would water be?' So according to Chalmers, doing this sort of conceptual analysis is an a priori enterprise.

The main point for doing conceptual analysis for Chalmers is to give reductive explanations of various phenomena. A reductive explanation of a phenomenon for Chalmers is an explanation of the phenomenon in terms of microphysics. Specifically, let *P* be the conjunction of microphysical truths about the world.[18] Then consider the material conditional:

(5) $P \rightarrow$ there is water

If this conditional is knowable a priori, then *there being water* has a reductive explanation according to Chalmers. For in that case, 'we show that there is a sort of transparent epistemic connection between the microphysical and macrophysical phenomena.'[19] Now in judging whether 5 is known a priori we must be appealing to the primary intension of

[18] For the purposes at hand, Chalmers actually wants to conjoin *P* with a 'that's all' clause that says that the world contains exactly what is implied by *P*, and some indexical information. See Chalmers and Jackson 2001 pp. 317–318. I ignore this for simplicity here.

[19] Chalmers and Jackson 2001 p. 351.

'water' in considering the consequent. We are asking whether, if the world turns out the way *P* says, there is water. So for Chalmers, the interest in the primary intension of terms, and hence conceptual analysis, is in considering whether a reductive explanation can be given of what the term applies to by way of considering whether a conditional like 5 is knowable a priori.

As to worries with Chalmers' view, first it is questionable whether terms like 'water' really have primary intensions that are knowable a priori. Since the primary intensions encode how a term comes to pick out what it does in a given world considered as actual, the claim that such intensions are knowable a priori amounts to the claim that we can know a priori how our terms came to pick out what they do. See Laurence and Margolis 2003 for discussion. Further, the claim that 5 must be knowable a priori for water to have a reductive explanation is controversial. See Block and Stalnaker 1999 and Chalmers and Jackson 2001 for discussion. For pro and con discussion of other issues regarding the Chalmers–Jackson approach to conceptual analysis, as well as Lewis' discussed above, see Braddon-Mitchell and Nola 2009.

Finally, though Chalmers doesn't think otherwise, it is worth noting that conceptual analysis as Chalmers understand it is does nothing to address the paradox of analysis. For consider the following sentences:

(6) For all x, x is an instance of knowledge iff x is a justified true belief.
(6a) For all x, x is an instance of knowledge iff x is an instance of knowledge.

On Chalmers' view, the primary and secondary intensions collapse for these sentences; and both sentences have the same intension. Thus, they are not assigned any different semantic value on a view like Chalmers' by means of which we could avoid the paradox of analysis. To repeat, that this is not surprising since Chalmers' notion of conceptual analysis was crafted for another purpose.

References

Ackerman, D. F. (1981), 'The Informativeness of Philosophical Analysis', *Midwest Studies in Philosophy* VI, 313–320.

Ackerman, D. F. (1986), 'Essential Properties and Philosophical Analysis', *Midwest Studies in Philosophy* XI, 305–313.

Block, Ned, and Robert Stalnaker, 1999, 'Conceptual Analysis, Dualism and the Explanatory Gap', *Philosophical Review* 108, 1–46.

Bradden-Mitchell, and R. Nola, 2009, *Conceptual Analysis and Philosophical Naturalism*, Cambridge MA, MIT Press.

Chalmers, David J. 1996, *The Conscious Mind: In Search of a Fundamental Theory*, New York, Oxford University Press.

Chalmers, David J., and Frank Jackson, 2001, 'Conceptual Analysis and Reductive Explanation', *Philosophical Review*, 110, 315–361.

Chisholm, R. M. and Potter, R. C. (1981): 'The Paradox of Analysis: A Solution', *Metaphilosophy* 12(1), 1–6.

Jackson, Frank, 1994, 'Armchair Metaphysics.' In *Philosophy in Mind*, ed. J. O'Leary-Hawthorne and M. Michael. Dordrecht, Kluwer, 23–42.

Jackson, Frank, 1998, *From Metaphysics to Ethics: A Defence of Conceptual Analysis,* Oxford, Oxford University Press.

King, Jeffrey C., 1998, 'What is a Philosophical Analysis?' *Philosophical Studies* 90, 155–179

King, Jeffrey C., 2007, *The Nature and Structure of Content*, Oxford, Oxford University Press.

King, Jeffrey C., 2009, 'Questions of Unity', *Proceedings of the Aristotelian Society, Vol. CIX, Part 3*, 257–277.

Laurence, Stephen and Eric Margolis, 2003, 'Concepts and Conceptual Analysis', *Philosophy and Phenomenological Research* 67, no. 2, 253–282.

Lewis, David, 1966, 'An Argument for the Identity Theory', *Journal of Philosophy* 63, 17–25. Reprinted in *Philosophical Papers Volume 1*, 1983, 99–107, New York, Oxford University Press. I use the pagination of the latter.

Lewis, David, 1970, 'How to Define Theoretical Terms', *Journal of Philosophy* 67, 427–446. Reprinted in *Philosophical Papers Volume 1*, 1983, 78–96, New York, Oxford University Press. I use the pagination of the latter.

Lewis, David, 1994, 'Reduction of Mind.' In *Companion to Philosophy of Mind*, ed. Sam Guttenplan, 412–31, Oxford, Blackwell Publishers. Reprinted in *Papers in Metaphysics and Epistemology*, 1999, 291–324, Cambridge, Cambridge University Press. I use the pagination of the latter.

Schilpp, Paul A., 1942, *The Philosophy of G.E. Moore*, Open Court Publishing Company.

Soames, Scott, 2004, *Reference and Description: The Case Against Two-Dimensionalism,* Princeton University Press.

Sosa, Ernest, 1983, 'Classical Analysis', *Journal of Philosophy* 80, no. 11, 695–710.

CHAPTER 15

MODELING

MICHAEL WEISBERG

1. Introduction

I live on a racially diverse block in South Philadelphia. A little more than half of my block is Caucasian, a little less than half is African American, and the rest of the people are of Asian or Latino descent. Let's imagine three things about my block and the city it is in: First, imagine that everyone else on the block values living in a diverse neighborhood. Second, let's imagine that there is some comfort threshold that everyone on the block has. If, say, less than 30% of the block was African American, the African Americans currently living here might feel uncomfortable and decide to move. Finally, let's imagine that the whole city has this preference structure.

What will happen to the city in the long run? Will the city gradually move towards more and more integrated blocks like mine (because people value diversity)? Will blocks be relatively integrated, but with some being 30% African American and some 30% Caucasian (because those are the "floor" thresholds)? Will the city become even more integrated? Or will it become more segregated (because people aren't comfortable being in a very small minority)?

I have a very hard time imagining what would happen in this scenario. When I have asked friends about it, many of them see the 30% threshold as especially salient and think we will find pockets of 30% Caucasians and 30% African Americans. But almost no one who comes up with the right answer, which economist Thomas Schelling discovered (1978) by constructing a model.

Schelling's original model was concrete, consisting of a chessboard, dimes, and nickels. The squares of the chessboard represented addresses in a city, dimes and nickels represented households consisting of people from two racial groups, which I will call *A* and *B*. The dimes and nickels were distributed randomly throughout the board.

Besides the individuals and their initially random spatial layout, the model also contained a utility function and a movement rule. The utility function said that each individual prefers that at least 30% of its neighbors be of the same type. So the *A*s want at least 30% of their neighbors to be *A*s and likewise for the *B*s. Schelling's neighborhoods were defined

as standard Moore neighborhoods, a set of nine adjacent grid elements. An agent standing on some grid element *e* can have anywhere from zero to eight neighbors in the adjoining elements.

The model is made dynamic by a simple movement rule. In each cycle of the model, its agents choose to either remain in place or move to a new location. When it is an agent's turn to make a decision, it determines whether its utility function is satisfied. If if is satisfied, the agent remains where it is. If it is not satisfied, then the agent moves to the nearest empty location. This sequence of decisions continues until all of the agents' utility functions are satisfied.

When the movement rule and utility function are implemented in Schelling's physical model, something very surprising happens: a cascade is observed which leads from integrated neighborhoods to highly segregated neighborhoods. In a modern computer implementation of this model on a 51 x 51 grid (shown in Figure 15.1), a preference for 30% like neighbors usually leads to agents having 70% like neighbors.

Schelling urged his readers to actually take out a chessboard and implement his model so they could see the model's dynamics unfold. What one sees in doing this, or reimplementing the model in a computer, is a cascade: agents that start out satisfied become unsatisfied when a neighbor leaves or a new one moves in. This leads to movement, which leads to more agents becoming unsatisfied. A small patch of dissatisfaction can result in widespread movement, and ultimately, segregation. While there are a few agent configurations that are integrated where every agent is satisfied, these states are very rare and nearly impossible to generate from random agent movement. Thus, Schelling's major result is that small preferences for similarity can lead to massive segregation. This result is quite robust across many changes to the model including different utility functions, different rules for updating, differing neighborhood sizes, and different spatial configurations (Muldoon, Smith, and Weisberg, 2012). Schelling's model does what my imagination couldn't do. It predicts that my integrated block in South Philadelphia is unstable in the long run. Although the model is idealized in many ways, it says that over time, my block is likely to become homogenous if the movement rules and utility function of the model are anything like the ones that real people have.

Although Schelling's model is a simple one, I think it very nicely illustrates how an idea can be sharpened using a model. This sharpening forces us to examine our explicit and implicit assumptions, and extends the reach of our imaginative capacities, allowing

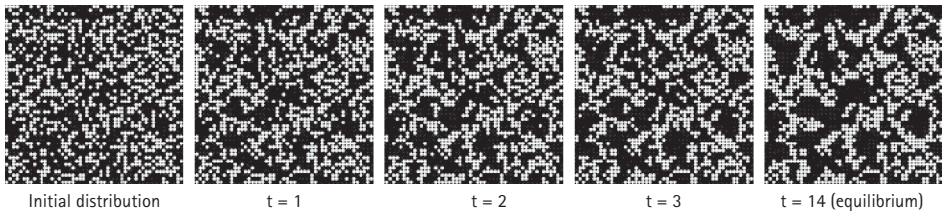


FIGURE 15.1 Computer simulation of Schelling's segregation model. On the left is shown a random distribution of the agent times. As time moves forward, large clusters of the two agent types form.

us to explain and predict complex phenomena that are difficult or impossible to gain a complete cognitive grasp of. This chapter explores the methodology of modeling, showing how it has been applied to philosophical questions and can continue to do so in the future.

2. What is Modeling?

Modeling is a form of surrogate reasoning, a practice in which one constructs and analyzes a model in order to learn, indirectly, about something else. Most commonly in both scientific and philosophical contexts, models are simpler than the real world target systems they represent and they are idealized relative to these targets.

Surrogate reasoning involves two steps: indirect representation of a target with a model and analysis of that model. One first constructs or acquires a model, and specifies the intended target of that model. This step does not involve extensive empirical or conceptual interrogation of a target and construction on the basis of inference from the properties of that target. Schelling didn't derive his model from a detailed description of Philadelphia or some other city. Instead, he asked himself about some of the essential properties of a city, and used those to create his model. So a model shouldn't be thought of as simply as a representation of a target, but rather an intermediary between the target and an analysis. This is why I call model-based reasoning "surrogate reasoning."

After constructing or acquiring the model, one subjects it to analysis. Techniques of analysis vary widely, and depend on both the type of the model and the question of interest. But typically one is interested in understanding the properties of various features of the model, and especially how some mechanistic features give rise to other behavioral features. Sometimes we try to give complete analyses, uncovering everything there is to know about the model. More often, analyzes are goal-directed, trying to answer specific questions. For example, Schelling's model can be used to answer questions about the tipping points or thresholds of segregation.

Modeling can be contrasted with *direct representation and analysis*. In this style of theorizing, one begins by representing a target system using what one knows about the target to generate an accurate representation. Although approximations and idealizations may enter the representation for pragmatic or epistemic reasons, the goal is to depict the features of some target system. If we wanted to study segregation by direct representation, we would look carefully at a time series of demographic information, such as data about each census tract. From this data, one could construct a representation of city migration patterns. One might also try to infer likely future patterns, or even the psychological motivations underlying them. It is possible that this would generate something like the Schelling model, but the procedure by which it was constructed would be very different. In direct representation, we analyze the system itself. In modeling, we study a constructed intermediary. The difference is one of practice and procedure, not necessarily the end product (Weisberg, 2007).

One of the virtues of modeling is that models are extremely flexible tools. They can be used to study a single target, a cluster of targets, a generalized target, or even targets known

not to exist. In philosophical contexts, models are rarely used to study a single target. Instead, they are most often used to answer *how possibly questions* (Dray, 1968; Resnik, 1991; Forber, 2010) or *what would happen if questions,* and hence usually have generalized systems as their targets. And sometimes, the targets of philosophical models are themselves hypothetical systems which do not exist, such as a perfectly just society, or a universe with only two particles.

3. Models

What kinds of things are models? This basic and central question has remained surprisingly controversial in the philosophical literature. Some philosophers, especially those who defend the semantic view of theories (e.g., Suppes, 1960; Suppe, 1989) argue that scientific models are the same kinds of things as logician's models. The motivating idea for this view is that theories should be language independent. Although we may describe theories with words, equations, and diagrams, they should not be tied to any of these descriptions. Proponents of the semantic view argue that the theory itself is a structure which satisfies such a description. A true theory, then, is a structure which is isomorphic to structures in nature. Models are thus a kind of mathematical structure.

More recent proponents of the semantic view, especially Bas van Fraassen (1980) and Elizabeth Lloyd (1994), argue that models are sets of trajectories in a state space. A state space is a set of points corresponding to the properties of a system. They are organized in such a way that each dimension of this space is an independent way that the state can vary. Trajectories through the space are time ordered sets of states that describe the temporal evolution of a model system. When there is some kind of match between the trajectories in the state space and trajectories corresponding to the target system, we have a model of the target system.

Another traditional view about models sees them as complements to mathematical theories, and hence not mathematical themselves. In this view, models are material analogies (Campbell, 1957; Hesse, 1966; for a dissenting view, see Duhem, 1906) they allow a scientist to develop an intuitive picture of a complex mathematical principle by comparison to something well-understood and concrete.

Hesse and others have emphasized that many important theoretical advances have been made when theorists understood that some property of one system was materially analogous to that of another. For example, the mathematics describing the propagation of light might be accompanied by an analogy comparing the propagation of light waves to the propagation of water waves. While we no longer think it is necessary for light to propagate though a physical medium, the analogy between light and water waves allowed James Clerk Maxwell to develop the equations describing light propagation.

Although most of the philosophers writing about the nature of theories today do not emphasize material analogies, this view has remained influential in the modeling literature in two ways. First, almost all philosophers of science accept that concrete models can do important scientific work. Watson and Crick used a material model of their proposed DNA structure to make inferences about base-pair hydrogen bonding (Watson, 2011). Walter

Newlyn and Bill Phillips constructed a hydraulic model to study how tax rates affect the British economy (Morgan, 2012). And the United States Army Corps of Engineers constructed a working tidal model of the San Francisco Bay and Delta Region in order to study what would happen if the Bay was dammed up (Weisberg, 2013).

A more controversial appeal to concrete systems can be found in a literature which asserts that all scientific models are fictional scenarios. Philosophers defending this view see all models, including mathematical models, as fictional scenarios that would be concrete if they were real. So on this view, Schelling's model isn't an abstract configuration of states and set of transition rules, but is actually an imaginary world: a neighborhood with people, preferences, and movement rules (Godfrey-Smith, 2006; see Weisberg, 2013 for a critique).

My own view of models is that they are composed of two parts: structure and interpretation. Like the critics of the semantic view, I think models cannot simply be mathematical objects. Bare structures stand in relations to target systems, but they often have too many of the wrong kinds of relations, and not enough of the right kind. However, mathematical, computational, or concrete structures, suitably interpreted, can stand in the right kinds of relations to represent features of targets. I call the relevant interpretations modelers' construals.

Construals provide an interpretation for the model's structure, set up relations of denotation between the model and real-world targets, and give criteria for evaluating the goodness-of-fit between a model and a target. They are composed of four parts: the *assignment*, the modeler's *intended scope*, *dynamical fidelity criteria*, and *representational fidelity criteria*. The assignment and scope determine the relationship between parts of the model and parts of the target system. The fidelity criteria are the standards theorists use to evaluate a model's ability to represent real phenomena.

Assignments are explicit specifications of how parts of real or imagined target systems are to be mapped onto parts of the model. This explicit coordination is especially important because although the parts of some models seem naturally to coordinate with parts of real-world phenomena, such as grid locations and addresses in Schelling's model, this is often not the case. For example, in a simple model of population growth, a population's growth is described by an exponential function. Nothing about this function suggests population size—it could just as easily signify a nuclear chain reaction. The theorist's assignment is what gives this function its meaning.

Assignments are often not made explicit in discussions of models, because communities of modelers have standard reading conventions for model descriptions. Where conventions are not explicit, are being violated, or where the modeler needs to be especially explicit, he or she will be forced to make the assignment explicit in discussions about the model.

Models inevitably have structure not present in the real-world phenomena they are being used to study. For example, Schelling's model has a perfectly regular grid and perfectly squared off edges. No actual city has these features. So are these features of the model intended to represent something about the target, or are they merely artifacts of the idealizations that went into constructing the model? A model's intended scope specifies the answer to this question, telling the theorist what parts of the model should be taken seriously.

The other aspects of a modeler's construal are fidelity criteria. While the assignment and scope describe how the target system is intended to be represented with the model, fidelity

criteria describe how similar the model must be to the world in order to be considered an adequate representation. I divide these criteria into two types: Dynamical fidelity criteria tell us how close the output of the model—the predictions it makes about the values of dependent variables given some independent variables—must be to the output of the real-world phenomenon. Representational fidelity criteria are more complex and give us standards for evaluating whether the structure of the model maps well onto the target system of interest. Typically, these criteria specify how closely the model's internal structure must match the causal structure of the real-world phenomenon to be considered an adequate representation.

For example, say that Schelling's model of segregation was targeted at the city of Philadelphia. One way to evaluate the model is with very high-fidelity criteria. If we did this, then the model's predicted equilibrium state, as well as the dynamics leading to that state, the utility functions of the agents, the movement rules, and so forth would be compared with the city's distribution of racial groups, looking for a very close match. Another way to evaluate the model is with a qualitative, not quantitative criterion. Yet another kind of fidelity criterion says that the model should be regarded as a how-possibly model, qualitatively matching the segregation patterns of the city, but with no expectation that the movement rules and utility functions were realistic.

Fully describing fidelity criteria requires an account of the model/target relation. If one thinks of models as (ideally) true descriptions of targets, fidelity criteria simply become an error tolerance, specifying how far one can deviate from truth. But when one has an account of the model/target relation that takes into account the highly idealized nature of many contemporary models, the situation is more complex.

Along with a concrete, mathematical, or computational structures, theorists' construals generate models. To say that a model is structure plus interpretation means that models are structures whose parts are interpreted via their assignments. They can potentially denote parts of a target as specified by the theorists' intended scope. And they are evaluated by the theorists' fidelity criteria. These four components of the construal constitute the theorists' interpretation of the model.

Whatever view about the nature of models is adopted, it is important to distinguish between models and their descriptions. Model descriptions specify models, and stand in many-many relationships to them. A single model might be described by words, equations, or diagrams. And any imprecision in a model description, including parameters left as dummy variables, will specify multiple models. Scientists often refer to equations as "models," but I think it is important to see equations as descriptions. Models' structures should be seen as independent of the way they are described.

4. Target Systems

Models are not compared directly to real phenomena, but to target systems, which are abstractions over these phenomena. The reason for this is that phenomena have many more properties than are represented in even the most realistic models. So when a modeler is ready to start comparing her model to the world, she constructs a target. She does so by

identifying a spatio-temporal region of interest and the contents of that region of interest. In scientific cases, the choice of target is driven by the research question of the scientist, specifically, which part of the empirical world is under investigation. Philosophical cases allow somewhat more latitude. Sometimes philosophers are interested in actual extant practices. Other times, they are interested in ideal scenarios such as conditions of perfect justice, or universes with minimal structure. Still other times, the goal of philosophical modeling can be the investigation of concepts, and there are no real or imagined targets for models.

Whatever the case, when a model is targeted at a real or imagined system, it represents only some parts of that system. Theorists must abstract away from the full richness of phenomena and aim their models at a set of features of a real-world phenomenon. For example, say I was interested in modeling a communication system. We might start by identifying the real-world phenomenon of people speaking English to one another. But this phenomenon is far too complex to capture in a model, so the modeler must decide which features to focus on. If the model was being constructed in philosophy of language, perhaps in order to investigate questions about intentionality, we might work with a very abstract target consisting of a set of symbols, states of the word, and transmission channels. However, a linguist would want to include many more details about the nature of language in her target, a communications engineer would include more about the transmission system, and so forth.

This example shows that the relationship between real-world phenomena and targets is one-to-many, which opens the door for a massive proliferation of target systems. Since there are so many different targets that can be generated from the same phenomenon, does anything go? Are there standards that govern the kinds of abstractions that theorists make?

Alkistis Elliott-Graves (ms) has argued that the answer to this question is no. Although many targets can be generated from one phenomenon, there are general norms for constructing appropriate targets. She argues that target system generation should be thought of as consisting of two conceptually distinct stages. Modelers partition the phenomenon into sets of features and then they abstract from these features in order to generate the target. Partitioning, she argues, is guided by the pragmatic norm of usefulness. The modeler should ask whether the relevant features for the topic of investigation get captured by the partition. Abstraction is more highly constrained by the norm of *aptness*, limited to what one can omit without distortion. Whether or not one accepts Elliott-Graves' account, it seems right to say that the enormous latitude of targets is not limitless. The flexibility it affords is positive, but the pragmatics of modeling impose limits.

Philosophical contexts, and, to be sure, some scientific ones, do not always require targets that are abstractions over real-world phenomena. More specifically, constructing and analyzing models of targets known not to exist (e.g. perpetual motion machines, time traveling bricks, or single particles alone in the universe) have played important roles in scientific and philosophical modeling. Sometimes, models are studied simply for their own sakes, without any target at all in mind. A good example of the latter category is Conway's Game of Life cellular automaton (Gardner, 1970). This model consists of an array of cells, which can each be in an alive state or a dead state. Transition rules determine how the states change, and these rules typically depend on the states of neighboring states.

One of the reasons we study models without targets is in order to help to sensitize our imagination so that we learn how to notice things we might have missed otherwise when looking at real targets. For example, Dennett discusses the interesting fact that when we begin thinking about the Game of Life, we start by describing a grid, cells, and the rules for each cell. But fairly soon we are talking about the patterns and apparent motion in the game.

> Note that there has been a distinct ontological shift as we move between levels; whereas at the physical level there is no motion, and the only individuals, cells, are defined by their fixed spatial location, at this design level we have the motion of persisting objects; it is one and the same glider that has moved southeast . . . Here is a warming-up exercise for what is to follow: should we say that there is real motion in the Life world, or only apparent motion? The flashing pixels on the computer screen are a paradigm case, after all, of what a psychologist would call apparent motion. Are there really gliders that move, or are there just patterns of cell state that move? And if we opt for the latter, should we say at least that these moving patterns are real?
>
> (Dennett, 1991)

Nevertheless, many of our most important cases of modeling are target-directed. A suitable target is chosen and the model is coordinated to that target by the modelers' construal. When that happens, what kinds of relations must a model stand in to its target?

5. Model/Target Relations

There are two types of accounts of model/target relations in the literature: *model-theoretic* accounts and *similarity* accounts. Model-theoretic accounts are the dominant view. Like other aspects of the modeling literature, they find their original home in discussions of the semantic view of theories. Such accounts typically posit that models must be *isomorphic* to their targets, although some proponents of the semantic view have weakened the requirement to *homomorphism* (Lloyd, 1994), or *partial isomorphism* (da Costa and French, 2003).

Isomorphism is a mapping between two sets that preserves structure and relations. Formally, an isomorphism is a bijective map between two sets such that the mapping function f and its inverse are both homomorphisms, structure-preserving maps between these two structures. This account of the model/target relation remains influential, but many philosophers have argued that it cannot appropriately deal with the relationship between idealized models and their targets (e.g. Hendry and Psillos, 2007).

As an alternative, Steven French and colleagues have offered the partial isomorphism account. Proponents of this account say that the model/target relation is tripartite, corresponding to the part of the model that is isomorphic to the target, the part that is not isomorphic to the target, and the part that is "left open" with respect to the target. A model is partially isomorphic to its target when a substructure of the model is isomorphic to a substructure of the target. Such an account can deal with some kinds of idealized models. For example, consider the idealization of elastic collisions that is associated with the ideal gas model. This idealization says that when two particles of the gas collide, the pair

maintains their combined kinetic energy after the collision. This is not true for molecular gases (hydrogen gas, oxygen gas, water vapor, etc.) because when they collide, some kinetic energy is transferred to the molecules' internal degrees of freedom (internal rotations and oscillations). However, the truth of this idealization is not required for many ideal gas model-based explanations and, in these cases, the idealized model can be confined to the non-isomorphic substructure without any loss.

But this will only handle idealization up to a point. In many cases, it is the idealized features of models themselves that are supposed to be representations of targets' features. For example, the Schelling model's idealized features, such as agents' utility functions and spatial distribution, are the very things that represent properties of real people and do the model's explanatory work. This has led some philosophers, including me, to look elsewhere for an account of model/target relations.

Similarity accounts posit that the model/target relation is one of similarity: a good model is similar to its target in certain respects and degrees (Hesse, 1966; Giere, 1988, Godfrey-Smith, 2006). Proponents of this type of account tend to defend their account along two lines. First, they argue that model-theoretic accounts do not have the resources to account for the relationship between the most common type of model and its target: idealized models relating to realistic targets. Second, they argue that modelers often talk and think about models as if they resemble their targets. This is taken as evidence for the nature of the relation.

There is a long tradition of skepticism about similarity. In "Natural Kinds," W. V. O. Quine argued that similarity was "logically repugnant" (Quine, 1969) because it couldn't be analyzed in terms of more basic notions. He also thought that mature sciences would dispense with similarity all together. In a more detailed discussion, Nelson Goodman (1972) agrees with Quine and adds another challenge. He argues that similarity is too promiscuous a relation to do any philosophical work. For any three objects, there will always be some respect in which two of the objects resemble each other more than the third. This, Goodman argues, shows that there can be no context-free similarity metric.

For many philosophers, this was the end of the matter. Positing similarity as the model/target relation was a dead end. Others took the criticism to be a constraint on a reasonable account of similarity; it must be a context-relative relation. For example, on Giere's (1988) account, a model must resemble its target in certain "respects and degrees." Cartwright (1983) argues that the relevant similarity between models and their targets is "behavioral similarity," meaning the similarity of the model's and the target's causal structures.

While this criticism of Goodman's was simply taken on board, Quine and Goodman also challenged proponents of similarity to give a reductive analysis, showing how some particular model and some particular target could be more similar to each other than other random models and targets. Much less has been written on this question, but some of my own work attempts to give such an analysis.

In *Simulation and Similarity: Using Models to Understand the World*, I argue that we can analyze model/target similarity in terms of weighted feature matching, and idea that has its origin in Amos Tversky's *contrast* account of similarity (Tversky, 1977; Tversky and Gati, 1978). The basic idea is that a model's similarity to its a target is a function of the features it shares and the features it doesn't share. Because Goodman is correct and there is

no general, context-free account of similarity, some of a model's and a target's features are weighted more heavily than others.

The account can be developed as follows: First, we begin a set of features Δ. This feature set can contain quantitative or qualitative predicated, including "is purple," "is to the left of ξ," "will rain with probability 0.9," and so forth. Further, for model *M* and target *T*, *m* is the set of features in Δ possessed by *M* and *t* is the set of features in Δ possessed by *T*. Modelers' fidelity criteria also implicitly provide a weighting function *f*(·), which is defined over the power set of Δ. The overall similarity of the model to the target is given by an equation of this form[1]:

$$s(m,t)=\frac{f(M\cap T)}{f(M\cap T)+f(M-T)+f(T-M)}$$

When modeling, it is customary to distinguish between the overall properties and patterns of a system (often called the "output" in computational and mathematical modeling) from the underlying mechanisms generating these properties. I will call the first set of properties *attributes*, and the second set *mechanisms*. It is important for many kinds of modeling, including philosophical modeling, that they be distinguished. The reason for this is that in some instances of modeling, we care far more about feature matching between one or the other type of feature.

As an example, we can return once again to Schelling's model. When the model comes to equilibrium, it contains racially segregated clusters driven by agents' utility functions and rules for movement. Attributes such as degrees of clustering are states of the model and mechanisms such as agents' movement rules are the transition rules of the model. Insofar as Schelling's model explains segregation in actual cities, then there has to be some relation between the model's attributes and the city's attributes. And there has to be some relation between the model's transition rules and the actual mechanisms that drive segregation in the city.

Now consider two other uses of Schelling's model. If it is used to ask a "what would happen if?" type of question, such as I opened this chapter with, all that is required is that the mechanisms of the model match the mechanisms in the scenario I wished to investigate. It might also be used to answer a how possibly question. Most American cities are highly segregated, even if their overall racial breakdown is mixed. What could possibly cause a racially mixed city to be segregated? Schelling's model offers one answer to this question. In evaluating a how possibly question, all that is required is that the attributes of the target (racially segregated neighborhoods in this case) are shared by the model.

Summing up the last few sections, we can say that most cases of modeling involve indirect representation, where a model is studied in order to learn about some real-world target. Models are interpreted structures which stand in relationships of similarity to target systems. Target systems are, in turn, parts of real-world phenomena.

[1] This is a much-simplified form of the similarity equations developed in Weisberg, 2013.

6. Asking Philosophical Questions with Models

Thus far, I have primarily spoken about modeling in a scientific context. I have done so both because modeling is most at home in the sciences, and because the philosophical literature about modeling is mostly about scientific modeling. In this section, I turn to several examples of the use of models in philosophy.

6.1 Fairness and the Social Contract

One well-known instance of philosophical modeling involves the application of game theory to foundational issues in political philosophy. Philosophers have long been interested in the question of why rational, egoistic agents would develop the sense of fairness and justice that seems to be the heart of stable political arrangements. Looking to the early modern tradition, Jean Hampton (1988) and David Gauthier (1986) showed that some of Hobbes' arguments could be reformulated as game-theoretic models. More recently, Brian Skyrms (1996) and his students have further developed these ideas and applied evolutionary game-theoretic models to question about the origins of our sense of fairness.

To take just one example from this rich literature, let's consider the game called Divide the Cake. In this game, two players are given a chocolate cake and they have to figure out a strategy to divide it before it spoils. Each player asks for a certain fraction, and as long as those fractions add up to 1 or less, then both get some cake.

As Skyrms points out, the intuitively correct answer is also the fairest one: both should ask for half the cake. But there is nothing special about this solution. All divisions that add up to a whole cake are *Nash equilibria*, meaning that neither player could be better off changing her strategy unilaterally if that division is employed.

Despite the infinite number of equilibria, we have a very strong sense that the fair division is a 50/50 split. How could that be? What Skyrms was able to show is that when a population of dividers repeatedly plays the game and modifies their strategies according to the ones that lead to the biggest payoffs in cake (formally, employing the replicator dynamics), then the fair strategy can evolve. Skyrms estimates that with nothing else added to the model, the 50/50 division evolves 62% of the time (1996). However, once *correlation*—interacting with some players more frequently than others—is added to the model, then the fair split evolves most of the time. Skyrms writes:

> In a finite population, in a finite time, where there is some random element in evolution, some reasonable amount of divisibility of the good and some correlation, we can say that it is likely that something close to share and share alike should evolve in dividing-the-cake situations. This is, perhaps, a beginning of an explanation of the origin of our concept of justice.

Thus, Skyrms is able to show that in repeated interactions with correlation under an evolutionary dynamic, which might be instantiated in cultural evolution as much as in biological evolution, fairness norms begin to establish themselves in the population. Skyrms himself primarily offers this as an explanation of the origin of these norms, a sort of genealogy of

morals. But some philosophers take this kind of modeling to have normative conclusions. Assuming that we are primarily self-interested, norms of reciprocity and fairness are justified by their good outcomes, as judged by an egoist.

6.2 Meaning and Signaling

A second case of philosophical modeling concerns the origin of meaning. In Lewis' *Convention* (1969), he introduces a game-theoretic analysis of convention and uses this analysis to investigate how communication can arise without a prior shared language or communication system.

In the two-agent version of the model, we imagine that the first agent (the sender) observes the world is in some state S_i for which the second agent (the receiver) ought to perform action R_i. The agents are cooperative such that for each i, if the world is in S_i the sender wants the receiver to perform R_i. In order to achieve this, for each observation, the sender sends a signal σ_j which is received by the receiver. The receiver's contingency plan specifies the action R_i that will be performed for each signal σ_j. Both sender and receiver aim to achieve a signaling system with the following structure:

$$\begin{aligned} &S_1 \Rightarrow \sigma_1 \Rightarrow R_1 \\ &S_2 \Rightarrow \sigma_2 \Rightarrow R_2 \\ &S_3 \Rightarrow \sigma_3 \Rightarrow R_3 \\ &\quad\vdots \end{aligned}$$

To illustrate this kind of signaling system, Lewis asks us to consider a signaling system that may have been established between the sexton of the Old North Church and Paul Revere. If the sexton saw the British army staying at home, he would hang no lanterns in the belfry. If they set out to attack by land, he would hang one lantern. And if they set out to attack by sea, then he would hang two lanterns. Revere would stay home if he saw no lantern, warn of a land attack if he saw one lantern, and warn of sea attack if he saw two lanterns. The connection between number of lanterns and type of attack is, of course, purely arbitrary. One lantern might have caused Revere to warn of a sea attack. All that matters is that there is coordination between the sexton and Revere such that the signal leads to the right outcome.

In the actual historical case, Revere and the sexton agreed on the code. But what if they hadn't? Could this signaling system evolve by itself? More generally, when two agents (or organisms) have a common interest in communicating, but no shared language, can such a system evolve? Skyrms (2010) investigated this question by adding learning (change within an organism's lifetime) and the replicator dynamic (evolution between the lifetime of organisms) to the Lewis signaling model.

To investigate learning, Skyrms coupled a Lewis signaling model and Herrnestein's matching law. On this probabilistic model of learning, probabilities for taking actions get updated according to the reward accumulated at previous times. When these dynamics are applied to the simplest case involving two signals and two actions, the signaling system (signal 1 leads to action 1, signal 2 leads to action 2, …) is very quickly learned. Skyrms also investigated this process from an evolutionary point of view. When the same type of setup

is allowed to evolve under the replicator dynamics, signaling systems are the only stable equilibria. Things get more complicated with greater numbers of signals and actions, but the overall lesson is the same.

6.3 The Division of Cognitive Labor

The final example that I will discuss comes from philosophy of science. There was a long tradition in philosophy of science that saw ideal scientists as impartial, cooperative, motivated by the truth, and always striving to do highly significant work. Although philosophers knew that real scientists could fall short of these ideals, they took this to be a more-or-less accurate description of scientists most of the time and the ideals that scientists should strive for. Scientific communities function better when they cooperated in order to find out the truth.

This picture has been called in to question by historians, sociologists, and philosophers. The classic source for such doubts is Thomas Kuhn's *The Structure of Scientific Revolutions* (1962). Kuhn argued that much of scientific inquiry took the form of "articulating the paradigm," a kind of incremental, puzzle-solving activity. He also argued that in times of scientific revolution, resolution of theoretical controversy had more in common with religious conversion experiences then with rational discourse. Others following in this tradition emphasized all the non-rational and even irrational qualities that characterize scientific behavior. Instead of priests in lab coats, powerful scientists look like mafia bosses, and their underlings like foot soldiers.

Let's say that the historians and sociologists are right, and that much about science seems less than rational. What are the epistemic consequences for the scientific enterprise? Does lack of communication, lack of epistemically pure motivation, and a focus on small-scale puzzles mean that we have to change our views about the authority and productivity and science? A number of philosophers have tried to address these questions by modeling, and the resultant literature is about the division of cognitive labor.

Among the best known philosophical models in this area are those offered by Philip Kitcher (1990) and Michael Strevens (2003). Kitcher and Strevens focus on the question of motivation: What happens when scientists have motives other than the truth? Specifically, what happens when individual scientists are motivated by getting credit for important discoveries (presumably because this leads to fame, higher ranked positions, and money), rather than simply learning about the truth.

Imagine that the scientific community has a particular scientific goal in mind—in his original scenario, Kitcher describes something like the race to find the structure of DNA. In order to reach that goal, there are n research approaches that might be taken. Each research approach has a success function, whose input is the number of scientists taking that approach and whose output is the probability that the problem will be solved using that approach.

With this type of model, we can ask two key questions: What is the optimal distribution of cognitive labor? And how will scientists' motivations, including non-epistemic motivations such as those for prestige and credit, lead to different allocations.

On the basis of such models, Kitcher and Strevens argue that classical epistemic norms will lead scientists to misallocate their cognitive labor. If a classically rational,

truth-seeking agent followed the procedure above, then he or she would join the project with the highest probability of success. But this isn't always what the scientific community as a whole wants to see happen. Maximizing the chance at success might involve distributing scientists across projects, not just to the projects with the best chance of success.

What if scientists were motivated by the accumulation of credit, the recognition that comes from being the first to make a discovery? To answer this question, let's assume that the scientific community adopts the *Priority Rule* (Merton, 1957), that most or all of the credit for a discovery goes to the scientist who makes the discovery first. In this case, scientists will want to take into account both the probability of success of the project and the probability that they will be the first one to complete the project. The first consideration pushes scientists towards the project with the overall highest probability of success, but the second consideration pushes scientists towards projects that have fewer scientists working on them.

To investigate this question more fully, Strevens (2003) models a representative agent who is poised to enter the field for the first time. If this agent can choose between n projects, and knows the current distribution of scientists to projects and the success functions of these projects, which one would she choose? Strevens shows that if the scientific community allocates credit according to the Priority Rule, then the community as a whole achieves the optimum division of cognitive labor. This, he argues, explains why the scientific community has adopted the Priority Rule, the rule that whoever discovers something first gets all the credit. So even if scientists strive for credit instead of truth, the scientific community is well functioning. Further, it might actually function better if scientists strive for credit than if they are only interested in the truth.

Another line of research in this area has looked at cooperation in science. Famous laboratories such as the wartime Los Alamos nuclear weapons laboratory (Rhodes, 1987) and MIT RADAR laboratory (Galison, 1997) planned ways for scientists to continuously integrate their findings and share ideas with one another. Technological innovations such as the Internet and rapid forms of electronic publication make this possible on a much wider and geographically distributed scale. According to classical norms of scientific inquiry, this is unambiguously good and should be encouraged. But it this really true? Might such communication lead to the propagation and fixation of errors as well as knowledge?

Epistemic network models allows us to investigate these questions (Zollman, 2007; see also Grimm et al., forthcoming). In such models, lines of communication between scientists are represented by network graphs, such the ones seen in Figure 15.2. Each node of these graphs represents a scientist and each edge a communication channel. By altering the connectivity of the graph, from the minimally connected cycle to the maximally connected complete graph, we can represent different types of scientific communication—from maximal to minimal.

How can this be used to investigate norms about communication? In a recent paper, Zollman considers networked scientists trying to decide what propositions are true of the world. Imagine that scientists are trying to determine whether the world is in state s_1 or s_2. They get information from their own experimentation and also from the other scientists they are connected to. On the basis of this information, they update their beliefs in a standard Bayesian way. When their beliefs reach some threshold (i.e. the probability is high enough), then they decide what to believe. The main independent variable in the model is connectedness of scientists.

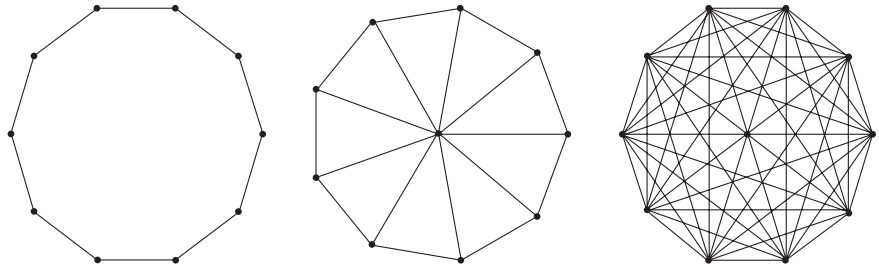


FIGURE 15.2 Three epistemic networks explored by Zollman. The nodes represent agents, and the edges represent lines of communication.

Images courtesy of Professor Zollman

Zollman's model generates two especially interesting results. First, scientists connected in a cycle converged to the truth more often than scientists connected in a wheel or in a fully connected graph. This suggests that careful limiting of information available to scientists may have certain advantages. Or to put it another way, less well-informed scientists might have an advantage over more well-informed ones, if the goal is to minimize error. However, when more communicative communities converge to the truth (or a falsehood), they do so more rapidly. For the ten-scientist communities Zollman studied, those on complete networks converged about five times faster than those in the cycle network. So one might conclude from this that too much communication is a bad thing. Scientific communities may be better off when they partially limit communication, to ensure that the wrong answers aren't locked in too quickly.

A final line of philosophical modeling in this area considers how scientific communities discover significant significant scientific truths. Epistemic landscape models (Weisberg and Muldoon, 2009) investigate the ways that scientists choose what kinds of problems to work on and the approaches they take in order to do so. They begin by postulating a set of approaches, narrow specifications of how a research topic is investigated. These approaches are then organized by their mutually independent properties. Each one of these properties is represented as a dimension in an epistemic landscape, whose points correspond to approaches. An additional dimension corresponds to the epistemic significance of the approach, giving a topography to the landscape where peaks correspond to the most highly significant approaches, as in Figure 15.3. However individual scientists are motivated, a socially optimal outcome would be one where the peaks are found and so are the many highly significant regions that are not peaks.

Scientists are represented in epistemic landscape models as agents who make strategic choices about what approaches to take. They get feedback from the landscape about the significance of the approaches they have taken, and have the possibility of communicating with other agents about the significance of the approaches that these other agents have taken. So an exploration strategy will be the rules an agent follows in determining which approaches to adopt in each cycle of the model. Should it keep the current approach, or move on? Should it take into account what others are doing? If so, how should this information be taken in to account.

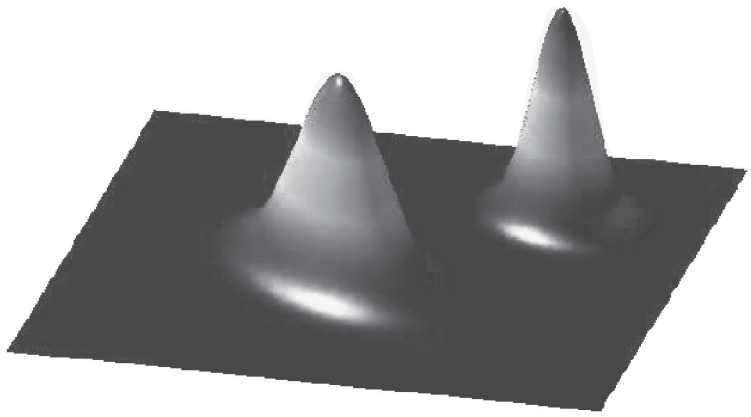


FIGURE 15.3 A low dimensional epistemic landscape investigated in Weisberg and Muldoon, 2009. The *x* and *y* dimensions correspond to aspects of the research approach and the *z* axis corresponds to degree of epistemic significance.

Weisberg and Muldoon, 2009

In one model, Muldoon and I investigated the last of these questions. We looked at two extreme ways that scientists can take account of what others are doing in their search for approaches of high significance. Followers think that the best way to find more significant truths about the world is to find the approach which has yielded the highest significance so far, and move in that direction. This is simulated in several steps. At the beginning of each cycle of the model, followers examine the patches in their Moore neighborhood, the eight approaches immediately adjacent to the one on which they are currently located. These patches correspond to the most conceptually similar approaches to the agent's current approach. Model agents then move to the previously explored approach of maximum significance in their Moore neighborhood, if such an approach is available. Like followers, mavericks pay attention to what others are doing, but they use this information differently. Instead of moving towards approaches yielding high significance, mavericks move away from explored territory.

When first working on this project, Muldoon and I expected the followers to do quite well because the strategy is essentially imitative. We predicted that the followers would help each other get to the "frontier" of unexplored knowledge, then they would spread out and discover what there as to discover. Mavericks, we predicted, would do less-well. Since they are always adopting new approaches, they never allow themselves to build on the knowledge of those that came before them, which seems to create a disadvantage if they are trying to find the peak of the epistemic landscape.

As it turns out, our imagination led us widely astray. When we take these ideas about maverick and follower strategies and implement them as models, we see a very different result. For ease of visualization, let's look at a simple, three-dimensional landscape: two dimensions correspond to ways that approaches can vary, a third corresponds to epistemic significance. At the beginning of a simulation, scientist-agents are placed in random, low-significance regions of the landscape. They are then allowed to implement the follower

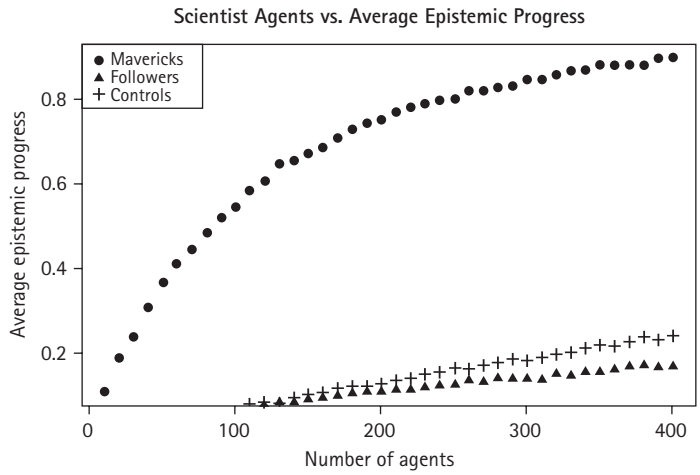


FIGURE 15.4 Epistemic progress of communities of agents of different types.

Weisberg and Muldoon, 2009

strategy or the maverick strategy. The results are striking: the mavericks massively outperform the followers.

Figure 15.4 shows the *epistemic progress* (fraction of significant approaches investigated) against the number of scientists of different types. As we can see from this graph, even small populations of mavericks massively outperform followers. So our intuition that followers would be effective was incorrect. In fact, we also were able to show that populations of followers do no better, and sometimes do worse, than populations that don't share any information. This result suggests that while the sharing information can be a good thing, it can also have unforeseen consequences. Sometimes the community is better off "spreading out" in epistemic space, not simply building on the best that came before.

I opened this section by saying that the traditional view of scientists was that they were impartial, cooperative, motivated by the truth, and always striving to do highly significant work. Kuhn and others called this picture into question, and drew radical epistemic conclusions from it. The philosophical modeling described in this section shows some of the ways that these less-than-ideal individual epistemic virtues may nevertheless be beneficial at a societal level.

7. Relationship between thought experiments and models

Throughout this chapter, I have often introduced the models I was discussing in the way one introduces thought experiments. I might introduce Schelling's model by saying, "Imagine a city where all the houses are arranged on a grid." This certainly sounds very similar to many philosophical thought experiments: "Suppose that I'm locked in a room and given a large batch of Chinese writing" (Searle, 1980). "Suppose you are the driver of a

trolley. The trolley rounds a bend, and there come into view ahead five track workmen, who have been repairing the track" (Thomson, 1985).

In this section, I want to think about this apparent similarity between thought experiments and simulations and ask a couple of questions: Are models and thought experiments the same kind of thing? Are models better than thought experiments? Should models replace thought experiments in philosophical theorizing?

7.1 Thought Experiments and Models

Although there is a large philosophical literature about thought experiments (e.g. Gendler, 2000a; Sorensen, 1992), there is little consensus about what kind of thing thought experiments are. Most philosophers accept that thought experiments are imaginary scenarios, and that is how I will understand them in this chapter. If thought experiments are imaginary scenarios contemplated in order to help us learn something about the world, then they look very much like models. The main difference is the role of imagination: I have argued that models are interpreted concrete, mathematical, or computational structures. Thought experiments, on the other hand, are products of imaginations. What is the relationship between these things?

One view is that all models are actually a kind of thought experiment. Godfrey-Smith defends what I call the *fictions' view*, arguing that even mathematical models are best understood as fictional scenarios. In discussing the ways that modelers talk and think about their models, he writes:

> I take at face value the fact that modelers often take themselves to be describing imaginary biological populations, imaginary neural networks, or imaginary economies. An imaginary population is something that, if it was real, would be a concrete flesh-and-blood population, not a mathematical object. (Godfrey-Smith, 2006)

On a more standard view of models that sees them as structures, not all models are thought experiments. There are many mathematical and computational structures that are difficult or even possible to imagine, so they can't be the same kind of things as thought experiments, unless one thinks that thought experiments do not literally have to be imagined. So on this view, the set of models is clearly much larger than the set of thought experiments. But the converse question remains: are all thought experiments models or proto-models?

It is tempting to simply say that thought experiments are models, where imaginative structures take the place of mathematical, computational, or concrete structures. A more refined way to link thought experiments to models is to say that the narrative parts of thought experiments are model descriptions for concrete or computational models. Without seeing the dynamics of Schelling's model play out on a computer screen, I can't imagine much about the scenario he describes. But here is a very simple Schelling-like thought experiment: There are 10 houses on a block and only four families, half Caucasian and half African American. Each family wants to have 50% of its neighbors be of the same race and, although they are initially placed randomly in houses, they can move freely in the block until they find a configuration that satisfies them. It is easy to imagine the equilibrium state of the model: four houses together alternating in racial makeup.

What was this thought experiment I was engaging in? There are clearly the same kinds of computational elements as Schelling's original model: a configuration of agents, a utility function, and movement rules. The main difference is that this case is sufficiently simple to have a good intuitive grip on, that could be checked with a computer program or chessboard. So in this case, the thought experiment is functioning in exactly the same way as a computational model.

Here is another kind of case. Gendler (2004) asks us to

> [t]hink about your next-door neighbor's living room, and ask yourself the following questions: If you painted its walls bright green, would that clash with the current carpet, or complement it? If you removed all its furniture, could four elephants fit comfortably inside? If you removed all but one of the elephants, would there be enough space to ride a bicycle without tipping as you turned?

What is happening in this case when we contemplate, along with Gendler, whether bright green would in fact clash with the carpet? Gendler argues that our judgments about these questions are made by creating the relevant mental image of green walls, carpets, elephants, and bicycles, and then forming a judgment by examining this image. She says that using a mental image to determine if elephants will fit is analogous to taking "a three-dimensional scale-model of the room, along with four similarly scaled plastic elephants … putting the elephants into the room, and seeing whether they fit" (1158). So in this case, we are using our imagination in exactly the same way that we would use a concrete model.

7.2 Advantages of Models

If thought experiments are models or proto-models, are there any advantages to being more formal and constructing full-blown models to replace thought experiments? Is it sufficient to rely on thought experiments, or should we take the construction of a fully explicit model as some kind of regulative ideal? In order to answer these questions, I think it is worth considering some of the advantages models have over thought experiments.

In a philosophical context, the main advantages of modeling over thought experiments are explicitness, reduced inter-philosopher variation, and the ability to deal with imaginative resistance. Creating and analyzing a model is a process that involves, among other things, forcing oneself to make all assumptions explicit. Our imaginations are very flexible and we can construct an imagined scenario from a very minimal script. But when one has to derive an equation, build something out of plastic, or write a computer program, this kind of vagueness is not allowable. Programs won't compile, equations cannot be derived, and plastic models will fall apart if details are left unspecified. One of the most common experiences reported by model builders is discovering a missing or hidden assumption in the course of modeling. This involves finding a resolution, or realizing that there are multiple avenues worthy of investigation.

A second major advantage is the reduction of inter-philosopher variation. There will always be philosophical disagreements about how to assess the results of thought experiments. This is especially true in normative domains, but also true in epistemology and metaphysics cases. However, sometimes our imaginations cannot even resolve the

phenomenon we are supposed to be judging, or we seem to think different things would happen. For example, here is a thought experiment suggested by philosopher Simon May in personal correspondence: Imagine we allowed and encouraged professors and students to carry firearms on campus. Would there be many fewer or many more campus shooting fatalities? I can imagine this scenario, and I even know what I think would happen. But I have no confidence that my imagination is able to resolve the scenario properly and, therefore, no confidence in my judgment of the case. A major advantage of modeling in such cases is a determinate answer about what would happen in such a case, given such-and-so assumptions.

Finally, there are many scenarios that we are simply incapable of imagining, a phenomenon often called imaginative resistance (Gendler, 2000b; Walton, 2006). Although morally repugnant cases of imaginative resistance have received the most attention in the literature, the complex counterfactuals, high-dimensional spaces, massive aggregation over agents, and atypical mental states that arise in philosophical thought experiments may also induce resistance. While some of these scenarios may also resist analysis by modeling, mathematical and computational models do not face the cognitive and memory limitations of humans. In such cases, a careful description of the setup and initial conditions may yield to computation, even if not to human imagination.

So should explicit modeling replace thought experiments in philosophical analysis? I think such a position is both too strong and premature. There are many cases where we can engage in a thought experiment, but really have no idea how we can create a model for the case. Moreover, if what I have said in this section of the chapter is true, thought experiments already have a modeling-like character, so it isn't obvious how much of an improved understanding we get by modeling. However, there are enough advantages to modeling that a reasonable norm might be as follows: if one can construct and analyze a philosophical model, then one should attempt it. Short of actually building a model, many of modeling's internal norms such as explicitness, publicness, cycling back and forth between constructing the model, analyzing the model, and revising the model, also make good norms for thought experiment analysis.

8. How to Get Started Modeling

In this final section, I will make a few brief comments about how to get started modeling. Unfortunately, it is hard to make such comments without being so general as to be unhelpful, or so specific as to only be relevant for a particular type of modeling. I will therefore divide these comments into two sections: the first about general principles of modeling and the second about agent-based modeling, a type of computational modeling particularly well suited to many philosophical questions.

8.1 Modeling Cycle

Modeling begins in much the same way thought experiments begin, by formulating a question to be answered and choosing a scenario to be investigated. Rather than the scenario

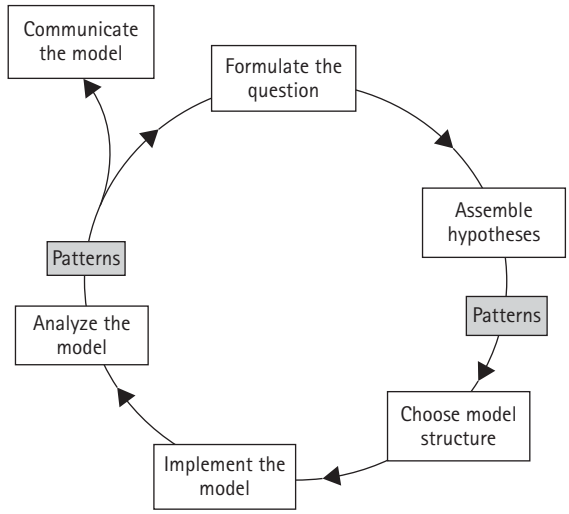


FIGURE 15.5 A depiction of the modeling cycle from Railsback and Grimm (2012).

Figure used with permission.

being something imagined, however, the scenario is what the model's interpreted structure will represent. In some of the recent literature on modeling methodology, this scenario is referred to as a hypothesis under investigation. Because models are stripped down versions of real-world scenarios, the modeler must pay special attention both to what gets "put in" a model. Grimm and Railsback (Grimm, et al., 2005; Grimm and Railsback, 2012) recommend that this be done in a pattern-oriented fashion, as depicted in Figure 15.5. One identifies patterns observed in some target systems and uses these to guide the construction of the model.

> Using observed patterns for model design directly ties the model's structure to the internal organization of the real system. We do so by asking: What observed patterns seem to characterize the system and its dynamics, and what variables and processes must be in the model so that these patterns could, in principle, emerge?
>
> (Grimm et al., 2005, 987)

How one goes about translating a scenario to a model depends on the type of model one wishes to build. If one wants to build an agent-based model, then one has to decide, among other things, who or what the agents are, what the agents are trying to accomplish, what their resources are, and what rules guide their decisions. If one wishes to make a game-theoretic model, then one has to identify a game that represents the scenario, the payoff structure of that game, and if a repeated game, the space of possible strategies. And so on for other types of models.

Across many model types, some general questions arise. For example, do the model's variables represent individuals or aggregates? Is time represented? Are the transition rules deterministic, probabilistic, or stochastic? Are these rules discrete or continuous with respect to time? Once answers to these questions and others are settled, one can choose a model structure and develop a construal.

Once the model is constructed, the process of analysis can begin. In general, there are two possible kinds of analysis here. If the modeler wishes to engage in complete analysis of the model, then he or she she will aim to determine:

1. the static and dynamic properties of the model
2. the allowable states of the model
3. the transitions between states allowed by the model
4. what initiates transitions between states
5. the dependence of states and transitions on one another.

I refer to this list as the total state of the model (Weisberg, 2013).

Complete analysis is usually associated with relatively simple mathematical models. In such cases, one can give analytical solutions which describe the models' behavior for every initial condition and every intermediary state. For more complex models, intensive computation, along with some approximation, can generate a complete or near-complete analysis of a model. But in many cases, complete analysis is too difficult to be practical, and not necessary. In such cases, modelers engage in goal-directed analysis, where they are investigating a specific set of properties or patterns of the model.

At the beginning of this chapter, I said that modeling was the process of indirect representation and analysis, and spent some time explaining how a model can represent a target in virtue of being similar to such a target. In order to "transfer" the results of an analysis of a model to a target, we need to know something about this similarity. Since models are almost never truthful representations of their targets, we are not looking for confirmation that the model is truthful. Rather, we are looking for validation, that the model resembles the target in certain respects, and then confirmation that the analytical results are confirmed in virtue of this validation.

8.2 Agent-Based Modeling

Much recent work in philosophical modeling, including many of the examples I have discussed, take an agent-based approach. Such models explicitly represent individuals, as opposed to aggregates with aggregate-level properties. For philosophers new to modeling, I suggest beginning with models of this type because of the availability of a straightforward, powerful, and free tool called Netlogo (Wilensky, 1999).

Netlogo is a high-level programming language, especially appropriate for creating simulations of social and natural phenomena that can be broken down into individuals or agents, which are called turtles in this framework. Although Netlogo is a powerful language that has been widely adopted by modelers across the natural and social sciences, it is very straightforward to learn due to its close association with a family of programming languages specifically designed for teaching. The original Logo language was one of the first pieces of educational software written in the late 1960s for the PDP-1.

The best way to get started with Netlogo is twofold: First, one should choose a few of its dozens of example models and explore them. Both the interface side and the programs themselves are well documented and explained. Much can be learned by modifying,

breaking, and fixing these existing models. Second, *Agent-Based and Individual-Based Modeling* (Railsback and Grimm, 2012) is an essential textbook for beginners. It combines helpful discussions on all aspects of agent-based modeling methodology with practical advice on programming in Netlogo, and it is accessible to complete beginners.

References

Campbell, N. R. (1957). *Foundations of Science*. New York: Dover.

Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford: Oxford University Press.

Da Costa, N. C. A., and French, S. (2003). *Science and Partial Truth*. Oxford: Oxford University Press.

Dennett, D. C. (1991). Real Patterns. *Journal of Philosophy*, 88(1), 27–51.

Dray, W. H. (1968). On Explaining How-Possibly. *The Monist*, 52(3), 390–407.

Duhem, P. (1906). *The Aim and Structure of Physical Theory*. Princeton: Princeton University Press.

Elliott-Graves, A. (ms) Target Systems and Their Role in Scientific Inquiry. (Unpublished doctoral dissertation). University of Pennsylvania

Forber, P. (2010). Confirmation and Explaining How Possible. *Studies in History and Philosophy of Science Part C*, 41(1), 32–40.

Galison, P. (1997). *Image and Logic: A Material Culture of Microphysics*. Chicago: University of Chicago Press.

Gardner, M. (1970). The Fantastic Combinations of John Conway's New Solitaire Game "Life". *Scientific American*, 223, 120–3.

Gauthier, D. P. (1986). *Morals by Agreement*. New York: Oxford University Press.

Gendler, T. (2000a) *Thought Experiment: On the Powers and Limits of Imaginary Cases*. New York: Garland Press.

Gendler, T. (2000b). The Puzzle of Imaginative Resistance. *The Journal of Philosophy*, 97, 55–81.

Gendler, T. (2004) Thought Experiments Rethought—and Reperceived. *Philosophy of Science*, 71: 1152–64.

Giere, R. N. (1988). *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.

Godfrey-Smith, P. (2006). The Strategy of Model Based Science. *Biology and Philosophy*, 21, 725–40.

Goodman, N. (1972). Seven Strictures on Similarity. In N. Goodman (ed.) *Problems and Projects*, pp. 23–32. Indianapolis: Bobbs-Merril.

Grim, Singer, Fisher, Bramson, Berger, Reade, Flocken, and Sales (forthcoming). Scientific Networks on Data Landscapes: Question Difficulty, Epistemic Success, and Convergence. *Episteme*.

Grimm, V., and Railsback, S. F. (2012). Pattern-Oriented Modelling: A "Multi-Scope" for Predictive Systems Ecology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367 (1586), 298–310.

Hampton, J. (1988). *Hobbes and the Social Contract* Tradition. Cambridge: Cambridge University Press.

Hendry, R., and Psillos, S. (2007). How to Do Things with Theories: An Interactive View of Language and Models in Science. In J. Brzeziński, A. Klawiter, T. A. Kuipers, K. Łastowski,

K. Paprzycka, and P. Przybysz (eds.), *The Courage of Doing Science: Essays Dedicated to Leszek Nowak*, pp. 59–115. Amsterdam: Rodopi.
Hesse, M. B. (1966). *Models and Analogies in Science*. South Bend: University of Notre Dame Press.
Kitcher, P. (1990). The Division of Cognitive Labor. *The Journal of Philosophy*, 87(1), 5–22.
Kuhn, T. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
Lewis, David. 1969. *Convention*. Cambridge: Harvard University Press.
Lloyd, E. A. (1994). *The Structure and Confirmation of Evolutionary Theory* (second edn.). Princeton: Princeton University Press.
Merton, R. (1957) Priorities in Scientific Discovery. *American Sociological Review*, 22, 635–59.
Morgan, M. S. (2012). *The World in the Model: How Economists Work and Think*. Cambridge: Cambridge University Press.
Muldoon, R., Smith, T., and Weisberg, M. (2012). Segregation That No One Seeks. *Philosophy of Science*, 79, 38–62.
Quine, W. V. O. (1969). Natural Kinds. In W. V. O. Quine (ed.), *Ontological Relativity and Other Essays*, pp. 26–68. New York: Columbia University Press.
Railsback, S. F., and Grimm, V. (2012). *Agent-Based and Individual-Based Modeling: A Practical Introduction*. Princeton: Princeton University Press.
Resnik, D. B. (1991). How-Possibly Explanations in Biology. *Acta Biotheoretica*, 39(2), 141–9.
Rhodes, R. (1987). *Making of the Atomic Bomb*. New York: Simon and Schuster.
Schelling, T. C. (1978). *Micromotives and Macrobehavior*. New York: Norton.
Searle, J. R. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3(3), 417–57.
Skyrms, B. (1996). *Evolution of the Social Contract*. Cambridge: Cambridge University Press.
Skryms, B. (2010). *Signals: Evolution, Learning, and Information*. Oxford: Oxford Univeristy Press.
Sorensen, R. A. (1992) *Thought Experiments*. Oxford: Oxford University Press.
Strevens, M. (2003). The Role of the Priority Rule in Science. *The Journal of Philosophy*, 100(2), 55–79.
Suppe, F. (1989). *The Semantic Conception of Theories and Scientific Realism*. Chicago: University of Illinois Press.
Suppes, P. (1960). A Comparison of the Meaning and Uses of Models in Mathematics and the Empirical Sciences. *Synthese*, 12(2–3), 287–300.
Thomson, J. J. (1985) The Trolley Problem. *The Yale Law Journal*, 94(6), 1395–415.
Tversky, A. (1977). Features of Similarity. *Psychological Review*, 84, 327–52.
Tversky, A., and Gati, I. (1978). Studies of Similarity. In E. Rosch and B. Lloyd (eds.), *Cognition and Categorization*, pp. 79–98. Hillsdale, N.J.: Erlbaum.
Van Fraassen, B. C. (1980). *The Scientific Image*. Oxford: Oxford University Press.
Walton, K. (2006). On the (so-called) Puzzle of Imaginative Resistance. In S. Nichols (ed.), *The Architecture of the Imagination*, pp. 137–48. Oxford: Oxford University Press.
Watson, J. D. (2011). *The Double Helix: A Personal Account of the Discovery of the Structure of DNA*. New York: Scribner.
Weisberg, M. (2007) Who is a Modeler? *British Journal for the Philosophy of Science*. 58, 207–33.
Weisberg, M. (2013). *Simulation and Similarity: Using Models to Understand the World*. New York: Oxford University Press.
Weisberg, M., and Muldoon, R. (2009). Epistemic Landscapes and the Division of Cognitive Labor. *Philosophy of Science*, 76(2), 225–52.

Wilensky, U. 1999. NetLogo. <http://ccl.northwestern.edu/netlogo/> (accessed September 18, 2015). Center for Connected Learning and Computer-Based Modeling, Northwestern University. Evanston, IL.

Zollman, K. J. (2007). The Communication Structure of Epistemic Communities. *Philosophy of Science*, 74(5), 574–87.

CHAPTER 16

INTUITIONS

JONATHAN M. WEINBERG

1. Introduction: Everyone Talks about Intuition, but No One Does Anything about It

One might have thought that the topic of *intuitions* is already an extremely well-covered topic within analytic philosophy today. However, once one distinguishes the methodological questions from the metaphysical and epistemological questions about intuition, one finds a vast wealth of work on the latter sorts of questions, and surprisingly little on the first. One can indeed find a great many papers on such questions as

- what sort of mental state are intuitions and/or what sort of mental activity is intuiting?
- what is the canonical logical form of a intuition?
- why are intuitions generally appropriate as sources of evidence?

Yet across all those very robust and philosophically engaging debates, one finds comparatively little attention to such questions as

- as a practical matter, what sorts of errors should one watch out for when appealing to intuitions?
- how might such errors be avoided, or conversely, how might one accidentally exacerbate any tendency we may have to such errors?

Disturbingly, one can find little to no explicit manualization of proper intuitive practice in print. In other disciplines, important methods are frequently the subjects of manuals, textbooks, and so on, and graduate programs in those disciplines may require one or more courses in methodology to provide the requisite training in the nuts-and-bolts of those research tools. In our own field of philosophy, we can contrast the fairly complete absence of any codified, trainable procedures of intuition deployment and their corresponding absence from our explicit professional training, with the existence of numerous textbooks in various formal methods, and the nearly universal requirement of at least

one graduate course in logic across PhD-granting programs in the English-speaking philosophical world.

Most of the methodological ink out there regarding intuitions has been spilled in rather all-or-nothing terms: *either* intuitions on the whole should be tossed out, prohibited from serious philosophical argumentation, *or* the way that we use intuitions now is pretty much fine as it stands. I suspect that a culprit here has been the specre of *skepticism*, and indeed a number of interesting skeptical arguments about intuition have proliferated and been considered in the literature (e.g. Pust, Cummins, Stich, Kornblith; Sosa, BonJour, Bealer, Williamson). But arguing with skeptics can unfortunately leave one with a methodologial bar rather too close to the ground. Skepticism is wrong about ordinary, unaided perception, for example—yet it is clearly of great methodological importance in the sciences that we do not simply rely on ordinary, unaided perception, not only in the sense that we avail ourselves of all sorts of perceptual aids and instruments, but more importantly, we've learned that we must be extremely proactive in shielding ourselves from such phenomena as experimenter bias. Just refuting a skeptic about perception is inadequate as a methodological guide in the sciences; so, too, refuting a skeptic about intuitions tells us all too little about how actually to go about deploying intuitions in our first-order philosophizing.

I suspect that we have collectively been assuming that addressing *philosophical* questions about intuitions (including skepticism) should be able to exhaust our interest qua *philosophers* in intuitions. But I fear that this plausible-sounding assumption is mistaken. There is no guarantee that any given discipline's preferred methodological tools are themselves adequate for addressing key questions that might be raised about those tools. Perhaps mathematics is a field that is methodologically autonomous in this way. But note that most sciences are not. Cognitive psychology, for example, relies at some points on results from neuroscience, and the tools of neuroscience today cannot be fully understood without a substantial grounding in modern physics, as one can tell from the names of such recent neuroscientific tools as "positron emission tomography" and "superconducting quantum interference device". And of course one needs to learn a fair amount of statistics to do contemporary psychology, even if that mathematics is not itself part of psychology. So we should not be surprised if it turns out that our methods as philosophers may face issues which specifically philosophical training is inadequate to address.

2. Maybe We Aren't Talking about How to Use Intuitions Because *We Don't Actually Use Them*?

Let me set the stage for discussing the methodology of intuitions by first considering a line of argument that Max Deutsch and Herman Cappelen have recently developed over the descriptive adequacy of what Cappelen calls "methodological rationalism," and their own preferred view which I will call "intuition nihilism" (henceforth IN). I will bootstrap from there, and from the inadequacies in both of those accounts, to my own descriptive take on intuition-deploying philosophical practice today.

Methodological rationalism (hereafter MR) affirms that philosophers do, as a matter of common practice, perhaps even as part of our "standard justificatory procedures" (Bealer), appeal to intuitions as a key source of evidence. Let's call that component of MR "intuitive practice." More than that, though, MR presents a substantive take on what these intuitions are, that it claims philosophical practice standardly appeals to. Cappelen extracts three main theses from MR regarding intuitions: that they have a "Rock" status as unshakably justifying without requiring justification themselves; that they have a distinctive phenomenology; and that they are grounded entirely in our conceptual competences. Cappelen investigates the primary philosophical literature, looking at a number of "test cases"—classic texts that a great many philosophers take to be paradigm instances of the methodology of intuitions in practice—and finds, quite simply, that none of these characteristics appear to be present in them. I think he is right about that, and this gives a strong reason to reject MR.

MR affirms intuitive practice (hereafter IP), and conjoins that further with a specific characterization of what intuitions are. IN uses that *latter* characterization, and the observed non-manifestation of MR's characteristics in the literature, to argue against the *former* claim that our practices do commonly use intuitions. One avenue of response for would-be defenders of IP would be to offer a different substantive account of what intuitions really are, and then demonstrate that intuitions under this rival characterization are, in fact, frequently deployed in the literature, and ideally they would show them to be present broadly across Cappelen's test cases. There is no shortage of candidates: intuitions are manifestations of our semantic competence, even without any special phenomenological features (Jackson; Jackman; Horgan and Henderson); they are a form of first-person access to the contents of our concepts (Goldman); they are deliverances of heuristic cognition (Weatherson); they are, at least some of them, products of our capacity for folk-psychological mentalizing (Nagel); they are rational insights derived from neo-Aristotelian mental contact with the universals themselves (BonJour); they are empircally derived implicit theories or conceptions (Devitt; Jenkins); they are simply a species of judgment in general (Williamson); they are the application of refined first-order inferential competencies (Ichikawa & Jarvis); they are the products of any of a potentially large hodgepodge of unconscious mechanisms (Weinberg et al. 2012).[1] That there are at present so many contenders for what intuitions really are, also suggests that probably none of them are right as a characterization of what philosophers on the whole take themselves to be appealing to. I am certain that we could apply Cappelen's procedure against each of these substantive accounts of intuition and show that none are obviously part of the workings of the practice itself.

We can, however, locate a fairly broad agreement across the profession as to where intuitions are supposed to be found, even while facing this broad dissensus as to what intuitions are supposed to be. In other words, in order to describe current philosophical practice adequately, we need to distinguish between the evidential role that intuitions are supposed to fill, and these various substantive accounts of what intuitions might be, such that they

[1] I should warn that I am lumping in some philosophers here who, while they take themselves to be offering substantive accounts of what underlies this philosophical practice, nonetheless would resist using the term "intuitions" in that account; see in particular Williamson and Ichikawa & Jarvis.

could fill them. While Cappelen concludes that all this business about intuitions is just a delusion of metaphilosophers, that inference is only legitimate if he can show that there is not really an intuition-shaped evidential hole in current practice, that such practice still requires something-or-other to fill.

So the strategy I will use to argue for IP, and thus contra IN (but without endorsing MR), will be to point out that hole—namely, the prima facie appearance that philosophers frequently ground some key substantive premises in their arguments, most famously verdicts on hypothetical cases, without any argumentation. (MR of course has an appealingly simple explanation to offer: at these spots, philosophers are appealing, perhaps tacitly, to intuitions as characterized in its substantive component.) While I think that Cappelen is right about Rock—and thus about MR—I worry that there is another, weaker and less presumptuous status, that might be more relevant to a wider discussion of philosophical practice once we move beyond the defeated MR. I'm going to call this status "Ground," in connoting still something that one stands on, and with positive methodological connotations, but perhaps in a squishier, less unyielding way than "Rock". Where Rock requires states to have a *special* status, unquestioningly convincing beyond the power of evidence to convince otherwise, Ground is just the status of being deployable as a default justifier whose status is not being secured by argumentation. MR, according to Cappelen, is committed to something like this (p. 105): "Intuitive judgments about cases are treated by philosophers as having a kind of epistemic status, which is different in kind from that of typical judgments based on perceptual input, memory, testimony, or those reached by inference from such judgments. The privileged status can be described as a kind of *default justification*" (original emphasis).

Even if we follow IN in rejecting the idea of there being a special status here for intuitions, distinct from the likes of perception and memory, we nonetheless might want to endorse the view that the intuitive evidential role is meant to involve exactly the same sort of status as the evidential roles of perception, memory, or testimony. This is thus an ordinary, not special, sort of status. This is still to be contrasted with the methodological status of the products of explicit inferences, whose acceptability depends at least in part on the status of the premises of those inferences.

Note that this is a *methodological* basicness, which need not be an ultimately *epistemological* one. Thus it is consistent with views that may legitimately appeal to these deliverances without further defense, at least in part because we *also* possess other sorts of evidence about our capacities here, which nonetheless need not be explicitly adverted to except under unusual circumstances. One might imagine something analogous to a coherentist sort of story about perception, where our justified theory about the reliability of our senses plays an important role in conferring justification on our perceptions, while of course coherentists can recognize that we can and do typically treat our perceptions as being default justified without having to think about or explicitly cite such a validating theory of perception.

The structure of the intuitive evidential role would also thus be fairly consonant with the roles in scientific methods configured by well-entrenched practices with the deliverances of measuring instruments. For example, one can typically simply report on temperatures as measured with a standard thermometer, taking those measurements to be default justified without requiring any further argument in defense of the claims that these were the readings of the instrument, or that the instrument is in fact trustworthy over this range.

Our background information about the trustworthiness of such instruments over such ranges may be a requisite part of a complete methodological story here, but such information's being broadly possessed in the background is what enables the instrument's deliverances to be appealed to in a methodologically basic way, that is, as Ground.

Though Cappelen makes a strong case that most of his "case studies" are a bad fit for Rock, many of those same examples seem actually to be pretty good examples of Ground. Here's Cappelen, objecting that Stewart Cohen's (1988) discussion of the lottery cases is palpably non-Rock: "The claim that philosophers rely unquestioningly on default justified, immediate reactions to these cases is simply unfounded: the moment the cases are presented, they are *questioned*, they *are not endorsed*, and they give rise to *puzzlement*. This response indicates that the Rock feature is absent" (p. 165; emphases original).

But the puzzlement that Cappelen observes to be attendant to Cohen's lottery cases, and which does seem to preclude the epistemically supercharged status of Rock, is nonetheless totally consistent with their being Ground. For claims that have default justification can still be questioned, unendorsed, and give rise to puzzlement. If I see something wildly unexpected—the department chair waltzing through the quad in a toga, say—then all three of those reactions may be highly appropriate, while still totally consistent with my perception having initially had a default justified status. Other illustrations of this idea can be found with optical illusions, M. C. Escher prints, and the like. And of course anomalous observations in the sciences are questioned, puzzled over, and ultimately endorsed only upon the conclusion of further investigation. But the scientists' undertaking of such questioning, puzzling, and investigation will often only make sense if we take it that their observation started with this kind of positive methodological status in the first place.

I have thus far argued that Ground is a highly common and ordinary sort of methodological status. But why think that it is on display in these sorts of allegedly paradigm instances of the appeal to intuitions, such as the case studies? The primary reason to think that current philosophical practice has an evidential role of this sort is the methodological basicness of case verdicts. Looking at a great many philosophical texts, especially many of those that are commonly taken as paradigm instances of the appeal to intuitions, it can seem that the verdicts about specific cases often bear a large amount of the argumentative load. (Indeed, it is often referred to as the "method of cases"; e.g. Johnston (1987), or in the subtitle of Nagel (2012), "A Defense of the Case Method in Epistemology".) And it can seem in particular that these case verdicts need to be secured without much at all by way of general argumentation—prima facie, they seem to be what we use in our arguments when more general arguments are inadequate to bring home our desired conclusions.

Nonetheless, Deutsch (2010) makes the case to the contrary, contending that philosophers do generally offer *arguments* for their claims in these papers, even in the most famous of (alleged) intuition-appeals, Gettier (1963). Cappelen follows this strategy as well in his (2012), proposing reinterpretations of many famous case-based philosophical texts so as to make the case verdicts depend on arguments. Yet many philosophical texts that look like arguments for a verdict on a given case, on further inspection are ones where the case verdict itself is doing the heavy lifting. There's really just no good way to understand the arguments in the vicinity of Lehrer's TrueTemp case and Thomson's violinist case (to take two more of Cappelen's case studies) without understanding them as relying on their specific verdicts about those cases. One cannot understand Lehrer as appealing to his internalism

as a premise to argue for his desired case verdict, for to do so would make it impossible to see why his text is seen as an important argument for the internalism itself. Similarly, what is at stake in Thomson's classic paper is whether or not one person's right to life is a sufficient ground, by itself, to impose substantial restrictions on the liberty of someone else; we thus must not read her as taking as a premise the insufficiency of such right-to-life claims, and arguing from there to the result about the violinist.

I think that we can see this dynamic well in the following passage from BonJour's classic "Externalist Theories of Empirical Knowledge":

> But it seems intuitively clear nevertheless that this is not a case of justified belief or of knowledge: Samantha is being thoroughly irrational and irresponsible in disregarding cogent evidence that the President is not in New York City on the basis of a clairvoyant power which she has no reason at all to think that she possesses; and this irrationality is not somehow canceled by the fact that she happens to be right. Thus, I submit, Samantha's irrationality and irresponsibility prevent her belief from being epistemically justified.
>
> This case and others like it suggest the need for a further condition to supplement Armstrong's original one: not only must it be true that there is a law-like connection between a person's belief and the state of affairs that makes it true, such that given the belief, the state of affairs cannot fail to obtain, but it must also be true that the person in question does not possess cogent reasons for thinking that the belief in question is false. For, as this case seems to show, the possession of such reasons renders the acceptance of the belief irrational in a way that cannot be overridden by a purely externalist justification. (60)

Note that the second half of the first quoted paragraph can make it *look* like BonJour is arguing for the conclusion that Samantha doesn't know, from premises about irrationality of a certain sort precluding knowledge. But the second paragraph reveals why this could not actually be a good reading of the first paragraph: the case verdict is what is doing the argumentative work, as a premise, in revealing to us the relevant claim about irrationality and knowledge, which is itself the conclusion in turn.

That the cases are leading the way is further substantiated by BonJour's own account of his methodology, when he writes earlier in that paper that the radical nature of externalism makes it hard to find much common ground at the level of general principles about knowledge or justification:

> The problem, however, is that this very radicalism has the effect of insulating the externalist from any very direct refutation: any attempt at such a refutation is almost certain to appeal to premises that a thoroughgoing externalist would not accept.
>
> My solution to this threatened impasse will be to proceed on an intuitive level as far as possible. By considering a series of examples, I shall attempt to exhibit as clearly as possible the fundamental intuition about epistemic rationality that externalism seems to violate. Although this intuition may not constitute a conclusive objection to the view, it is enough, I believe, to shift the burden of proof decisively to the externalist. (Bonjour, 56)

Just to be clear, I am not relying here on BonJour's use of "intuition", but rather on his professing to be using a case-first procedure. This should substantially undermine any attempt to read him as offering arguments from the epistemic generalities to the cases. He needs the case verdicts themselves to be doing the argumentative work. And I do not think BonJour is being idiosyncratic here: my read of the literature is that it is very common,

when philosophers deploy arguments with case-verdicts in them, that the case verdicts are methodologically basic in the manner I have suggested here.[2]

As further evidence, one can find numerous points in the literature where the author themselves seems to say that their capacity to offer any further defense of a key premise has run aground. See, for example, Fodor (1997), "Special Sciences: Still Autonomous After All These Years":

> As with most of the metaphysical claims one comes across these days, the one that I just made relies for its warrant on a blatant appeal to modal intuitions. But I think the modal intuitions that I'm mongering are pretty clearly the right ones to have. If you don't share mine, perhaps you need to have yours looked at. (154)

Another example from David Lewis (1996), "Elusive Knowledge":

> I started with a puzzle: how can it be, when his conclusion is so silly, that the sceptic's argument is so irresistible? My Rule of Attention, and the version of the proviso that made that Rule trivial, were built to explain how the sceptic manages to sway us—why his argument seems irresistible, however temporarily. If you continue to find it eminently resistible in all contexts, you have no need of any such explanation. We just disagree about the explanandum phenomenon. (561)

Many more such examples can be found in the literature; I offer these as exemplary, and from philosophers working at the very top of the profession.

To recap, then, there are some phenomena that IN does not have a very good account of: in particular, there are many places, including in many of the test cases, where philosophers use premises that are surely not Rock, but which do seem to have the methodological status of Ground, and where there are no good ways available to explicate their having that Ground status. And that's all IP needs.

Over and over again in the test cases, we find claims that patently lack any evidence of the special glowy phenomenology, or any attempt to base the relevant claims on purely conceptual truths. This is a clear lesson that Cappelen teaches us. Yet we also, over and over again, see the test cases as being inexplicably Ground. They are not, in fact, argued for from independently convincing and available premises—at least, not in a way that does not build on their status as Ground in the first place. Nor is there any ready-to-hand generally available evidence to offer on their behalf, which can be presupposed to be antecedently shared, of a sort that we can find intelligible in terms of observation or testimony from experts, or for that matter, from things like the deliverance of well-calibrated instruments, the products of scientific consensus, and so on. Philosophers seem on the whole fairly happy to use them as evidence, even in the absence of any such further evidence on their behalf. The profession seems to be in a state of substantial consensus, then, that we generally have some sort of cognitive grasp or other on these claims. And this consensus also extends pretty well to the basic contours of this capacity, for example, that it includes, but is not at all limited to, a pretty open-ended range of hypothetical cases with stipulated facts, including perhaps nomologically impossible ones. What the profession seems also to be in a state of

[2] See my 2014 for further engagement of Cappelen's arguments on this matter. See also Weatherson (2014); Chalmers (2014).

substantial dissensus about, however, is just what this grasp ultimately amounts to, as illustrated by the many-hued panoply of extant accounts of intuitions listed earlier.

One fact about this state of affairs that I find especially noteworthy is that even philosophers who operate with radically different professed accounts about what makes the appeal to intuitions legitimate, are nonetheless perfectly happy to share cases back and forth with each other, without batting an eye—and without drawing any methodological consequences from their meta-level disagreement to their first-order debates. BonJour and Goldman, for example, moot the same cases back and forth with each other, but the former takes himself to be an old-school high-church rationalist, and the latter takes himself sometimes to be doing a bit of first-person cognitive science concerning the structure of his internal neural representations.

Overall, then, though philosophers are prepared to acknowledge a bit of concern over this lack of an established validating story for these frequently deployed premises, nonetheless they do not seem prepared on the whole to lose *too* much sleep over it. For the philosophical community, or at least its relevant sub-communities, it remains a philosophical project for scrutiny and debate, and the method itself need not wait upon any such validation, and we can continue more or less smoothly in taking the relevant propositions as having the Ground status.

My theory of this range of philosophical practices, then, I will call *Protean Crypto-Rationalism*. Protean, because it can take many forms, even across a direct volley between authors in the literature, and seems also not to require any particular form at all. I will tag it with the label of rationalism, because philosophers seem to me to be presupposing something that rationalists presuppose, namely, that something about the human mind is a source of evidence about hypothetical cases and so on, over and above what is delivered from the senses. We take ourselves to have cognitive command of the truths about these sorts of cases, even though such facts are very far from the sort of thing we can simply observe. But it is a crypto-rationalism, because practically no particulars about the functioning of this capacity plays any standard role in the literature, and is consistent with views about the capacity that would not count as rationalist by the traditional criterion of being devoid of perceptual input (e.g. Williamson and Devitt).

Philosophers generally think that intuitions work—or at least that something does the sort of work for us that we use the term "intuitions" to cover– but without any sort of established account of why they work. I am thus endorsing IP, but denying that any particular substantive account of intuitions plays a role in that practice, in order to accommodate this observed dissensus. Now, IN can somewhat explain this dissensus about what intuitions are and how they work: since there aren't any intuitions, we should expect it to be hard to come to much agreement about them! But, in addition to facing the arguments I offer above about the primacy of cases, IN has trouble explaining why the profession on the whole—and not just a few wild-eyed metaphilosophers—has been so badly misled here. Why haven't more people noticed that they don't need to offer any sort of messy validation story, but instead have these other arguments and evidence available for these propositions, of the sort that IN appeals to? This is a bit mysterious on IN, but unproblematic for PCR.

PCR also does a better job of explaining the central importance of the cases, in philosophical disputes and in the discussion of how those disputes go. It is often unclear in IN's treatment of case studies, just why the author routed their argument through the

cases—especially such odd-sounding cases—instead of going after their target more directly. It is also unclear why, according to IN, it should be that we so often celebrate and transmit the cases from these classic articles, while simplifying or even leaving out altogether the (putative) arguments on behalf of those cases. (Cappelen notes that those who go on to talk about the cases often, by his lights, distort what was originally going on in the texts. But the widespread fact of this alleged distortion would itself be an explanandum, and a sign that at least those later philosophers were doing something different than what Cappelen takes these classic authors to be doing.)

Finally, if we do not acknowledge the existence of this evidential role for intuitions to secure such case verdicts as basic, then it is also hard to explain a lot of the back-and-forth about the cases that one observes in the literature. When one is faced with an unwanted case verdict, it is generally unacceptable simply to dismiss it—one might suggest that it is a "spoils to the victor" case, for example, or explain why this is an appropriate spot for the biting of a bullet. But what one pretty much never sees is anyone rewinding the case to a putative evidential source of the sort that IN would require for it, and trying to undermine that argument. For example, reliabilists felt compelled to reply to the Truetemp case, even though the (putative) premises in its defense were simply not ones that they would endorse. Relatedly, we can consider the phenomenon of the fairly common practice in which philosophers engage in something we are inclined to call "explaining away an intuition". PCR offers a very good explanation both of the appeal of this argumentative practice—and of why it generally doesn't seem to change anyone's mind. On the one hand, it is appealing, because when one's theory is inconsistent with p, and p has Ground, then one may want to offer defeaters for p, so it will lose that Ground status. On the other hand, as a strategy it rarely meets with offensive success, because the lack of much consensus as to how our intuitive capacities work means that it is very hard to offer a fully persuasive explaining away. They do seem to me to work reasonably well on defense, though, as a move to assuage one's own fears about an unwanted intuition.[3]

3. The Currently Sparse Epistemic Profile of the Appeal to Intuition

A key component of any piece of methodology is what I call its *epistemic profile* (Weinberg (2015): an account not just of its baseline reliability, but also of the particular sorts of errors it may be prone to. For example, the epistemic profile of a pocket compass would not just tell us what percentage of the time it is likely to point towards the north, but also detail its susceptibility to the nearby presence of large iron deposits, MRI machines, and the like.

According to PCR, the practice of the appeal to intuition does have a fairly agreed-upon shape, even if it lacks much agreement about what intuitions actually are or why they should be legitimately appealed to; accordingly, we should expect to be able to extract some, albeit limited, pieces of information that would go towards an epistemic profile of that practice.

[3] See also Ichikawa and Jarvis (2010) for further exploration of this facet of our practices with appeals to intuition.

I will briefly summarize here a few standard pieces of philosophical folk wisdom about the basic shape of the evidential role for intuition, and moreover, at least a few ways in which intuition can go astray. I am focusing here specifically on places where there is at least some methodological guidance offered, and as such I am leaving out a great deal of philosophical rumination on intuitions that does not offer much in the way of such guidance, for example, debates about whether intuitions are a *sui generis* mental state or a disposition to belief, or about how best to understand the underlying logical form of thought-experiment reports (see Williamson, Ichikawa, Malmgren).[4]

First, let me list four key aspects of the basic shape of the practice.

The primacy of cases: First, as noted in section 2, the practice licenses verdicts about particular cases. Now, it's certainly true that sometimes general principles are also offered as products of intuition, such as key premises in the sorites, or the universal prohibition on circular arguments. But one is hard-pressed to glean any clear sense of the methodological norms governing such appeals from the literature, so I will not pursue them further here. I do have the impression, though, that they are generally held as less secure than case verdicts are: when a case presents an intuitive counterexample to a principle, one can almost always reject the principle in favor of some appropriately weakened version. (The exception to all of this, I should also add, are the role of intuitions in well-regulated formal investigation, such as proofs in first-order logic.)

Flexibility of report format: These verdicts can be reported either as a first-order result (this case is not a case of knowledge), or metalinguistically (we would describe this case with the term "knowledge") or conceptually (this case does not fall under the concept KNOWLEDGE). Indeed an author who prefers one such format can engage in conversation with another author using a different format without any tension.

Freedom of stipulation: Our freedom to stipulate the particulars of the case to be considered—though obviously not including the verdict itself!—is restricted only by the limits of human short-term memory, and perhaps any bounds imposed by imaginative resistance. They are pretty much unconstrained by concerns about nomological possibility. Cases are typically hypothetical ones, though there is nothing in the practice that precludes considering actual cases as well (Williamson).

Interpretation-hungry: While the basic verdict itself on a case—for example, does the agent know, or not?—is part and parcel of the intuitive deliverance, more subtle aspects of that verdict may not be. In particular, we may not have the same degree of access to its precise modal force—counterfactually true, merely possible, or necessary, and if necessary, nomologically, metaphysically, conceptually, or logically? And we may not have any sense at all as to what factors were responsible for the verdict going the way it did. "No doubt, intuitions deserve respect … [but] informants, oneself included, can be quite awful at saying what it is that drives their intuitions; sometimes it's just a fragment of underdone potato. This holds all the way from chicken sexing to judgements of grammaticality and modality. Good Quinean that I am, I think it is *always* up for grabs what an intuition

[4] Let me emphasize again that I do not take these debates to be philosophically bad ones—in my view there is a lot of good philosophy there, but it just hasn't been directed at methodological matters. While it is surely possible at least in principle that such debates could yield practical implications for the conduct of inquiry, I don't think that any have yet been found.

is an intuition of" (Fodor 1998, pp. 86–7; emphasis original). And that is why the intuitive verdict of a case, while methodologically basic, will very often occasion a great deal of downstream contemplation and argument, as the philosopher tries to suss out just what it is about a set of often closely similar cases that rendered nonetheless disparate verdicts on them—the source of much of that *puzzling* about cases we have noted in the discussion of Cappelen.

In addition to these standard features of the practice, there are several other good candidates for at least *near*-consensus norms on our intuitive practices, but for which one can find at least a number of high-profile dissenters. For example, are intuitions to be preferred when they have been subjected to explicit deliberation, or would such conscious cognitive effort in fact pose a potential distraction from our 'real' intuitions? (see Talbot (2013) for a defense of the latter view). Is the felt strength of an intuition an important indicator of its trustworthiness? The following example from Davis (2007) exemplifies this proposed norm, as part of his argument for his own invariantist semantics for "knows" over contextualist rivals: "Neither semantic theory accounts for all of our semantic intuitions. The invariantist is forced to conclude that one of these intuitions of truth is incorrect. The contextualist must give up the even stronger intuition that the knowledge claims in the A and the B contexts are incompatible" (430). But other philosophers have suggested that we should be hestitant to put too much weight on considerations of felt strength (Williamson (2004); Saul (2007)). There is also ongoing debate as to whether intuitions about unusual or esoteric cases should be significantly discounted, in comparison with intuitions about more ordinary sorts of cases. (Fodor (1964); Kitcher (1978); Jackson; Williamson; Cappelen; Alexander & Weinberg; Jarvis & Ichikawa; McKenna).

In addition to these basic contours, the current epistemic profile of intuitions includes several sources of error for intuitions featured in at least the informal methodological lore of the profession, and sometimes they have merited explicit methodological treatment.

Misconstruals: One fairly tractable source of error involves misconstruing or misunderstanding the target scenario in some way. For example, undergraduates often fail to understand the disjunctive weakening inferential steps in Gettier's original cases, so they fail to appreciate how the beliefs of the agents in those cases are meant to be justified. These sorts of mistakes are usually weeded out by further reflection or deliberation, especially when the authors of the original cases are available. Philosophers are generally sensitive to the difficulties of entertaining complex cases carefully, and thus one frequently finds authors glossing and unpacking their cases in some further detail, in order to preempt such misunderstandings. Of more recent importance is the possibility of improper *filling in* of cases, since even fairly detailed scenarios will presuppose that the reader or listener round out the case with a huge range of unstated stipulations, very much along the lines of a reader or listener to a work of fiction (Ichikawa & Jarvis). The possibility of a mistaken filling-in may become particularly salient should the intuitions of non-expert subjects be considered, as with some work in experimental philosophy (Sosa (2009); Weinberg (2008)).

The literature also warns against a number of more subtle and less easily uncovered confusions.

Modal confusions: Kripke (1980, pp. 103–4) warns against another kind of error, in which we mischaracterize the precise modal content of our intuitions upon considering a case. In particular, Kripke warns that epistemic modality may be more permissive than fully

metaphysical modality, and thus we might intuit that, even while water is in fact H_2O, nonetheless it is possible that it be something else.

Bealer (1996) presents a Kripkean theory of error for modal intuitions, usefully with a recommendation for how to remove such errors. Using what he calls a *rephrasal strategy*, he recommends that one attempt to rephrase a putatively errant intuition that *it is possible that A*—say, that water could have not been H_2O—into something like the form *it is possible that a population of speakers in an epistemic situation qualitatively identical to ours would make a true statement by asserting "A", but with appropriately different referents in "A"* (pp. 133–4).[5]

Bealer comes later to reject rephrasal strategies in his (2004), for Burgean reasons, and he puts forward a different methodological strategy to discern where we may have mistaken a conflict between intuitions that really are divergent in their modal force: instead of reporting intuitions in terms of what is "possible", we should rather report them in terms of what is "contingent". Although "possible" admits of both epistemic and metaphysical readings, "contingent" is univocally metaphysical. So, if we will all report the intuition that "it is not contingent that water is H_2O", then there will be no difficulties here from a confusion between intuitions of epistemic and metaphyscial possibility.

Yablo (1993) identifies one source of error in terms of incorrect beliefs about either empirical matters, or about key metaphysical principles. Bealer (2002) extends that account with the further suggestion that one may suffer from a "local misunderstanding" of one's own concepts, but where such misunderstandings are "correctable *on [one's] own* (without the aid of any auxiliary empirical information) using the a priori … process, specifically, by careful examination of further cases" (p. 34; emphasis original).

More recently, Yablo seems to endorse a fairly open-ended understanding of possible modal errors. In his (2006), he rejects the following principle as an over-optimistic take on modal error:

> (O) Carefully handled, conceivability evidence can be trusted, for if impossible E seems possible, then something else F is possible, such that we mistake the possibility of F for that of E.

I would suggest that principle (O) would invite us, when concerned about the status of some putative intuition of possibility, to consider other candidate possibilia that might be responsible for this mistake. Since that set of potential doppelpossibilia would be fairly small, (O) would indeed give us cause for some optimism here.

Yablo ultimately can only endorse the somewhat less optimistic, but in no way skeptical:

> (O') Carefully handled, conceivability evidence can be trusted, for when impossible E seems possible, that will generally be because of distorting factors that we can discover and control for.

We are not, unfortunately, given any general recipe for such discovery and control. Yablo does not make it clear just what procedures we could adopt in order to be confident that we are handling our conceivability evidence sufficiently "carefully".

[5] There is also a rephrasal strategy involving mistaking a rigid designator for a nonrigid one, which Bealer offers several arguments against in his (2004). I will not pursue them further here.

Pragmatics/semantics confusion: Another kind of confusion that is part of our error profile of intuitions is the distinction between intuitively rejecting a sentence because we take it to be false, and intuitively rejecting it because it is not so much wrong as infelicitous. Of especial methodological use here is the test of cancellability, introduced by Grice (1975): if what we are finding intuitively off about a proposed sentence is that it violates a norm of conversational implicature, then we should be able to deny that implication explicitly, and thereby render the sentence acceptable. For example, in a situation where one has 10 philosophers in a room, each of whom is a metaphysician, is it correct to say "Some philosophers in the room are metaphysicians"? Or does that statement actually entail that at least some of them are not metaphysicians? One tests this by considering a statement like "Some of the philosophers in the room are metaphysicians—indeed, all of them were." That statement is intuitively kosher, and is thus evidence that the initial negative reaction was not semantic in nature. For an example of a test that goes the other way, consider the unacceptability of "The number of philosophers in the room was odd—indeed, it was divisible by two."

"Tin ear": The possibility of the "tin ear", in which a philospher is simply not adept at discerning how we would most naturally describe a situation, or at how best to map their own thoughts into language, is clearly part of at least the methodological folklore of the profession. For example, Buckwalter & Stich (2014) attribute to Ned Block the observation that "not all that long ago it was a common practice for philosophers to dismiss people who didn't share their intuitions by saying that they have 'a tin ear'." (n36) And Williamson claims that "philosophers with a tin ear for natural language sometimes seem to misarticulate their own strong intuitions, using forms of words that do not express what they really want to say" (p. 121). And one can find various appeals here and there to the general possibility of philosophers having this difficulty (e.g. Putnam (1990), p. 64).

But what has proved devilishly hard to find, despite the labors of a tedious afternoon on Google Scholar, are any actual *token instances* in print of one author lobbing such an accusation against another with regard to a specific case. I did find a recent instance in the *The London Review of Books* in 2012, in which Stephen Mulhall accuses Jeff McMahon of generally having a tin ear about the ethical significance of mortality, in his review of *The Ethics of Killing*. One rare instance of a specific accusation was in, of all places, *The Annals of the Association of American Geographers*, which does have the distinctly philosophical title "The Manipulation of Ordinary Language", by one Richard Symanski: his critic, David Sopher, accuses him of "undertaking to make delicate observations with a patently defective instrument, his own tin ear for the English language" (625), but even there it is about detecting hyperbole, and not a specific application to another philosopher's (or geologist's) appeal to an intuition about a case.

Not only can I not find much evidence of this maneuver in print, but accordingly I can locate very little guidance about how we are to avoid *tinnitus philosophicus*, or treat it when it has been diagnosed. It is not clear what should happen next, were two philosophers to have divergent intuitions and each accuse the other of this malady. I would suggest, though, that we might excavate some methodological ideas from J. L. Austin's ordinary language approach, as articulated, for example, his "A Plea for Excuses" (1956–7). Austin's suggestions there include: (i) preferring areas of language where ordinary practices are rich and well-established, which is part of why he thinks excuses make a good target (one might think this a reason to give greater weight to, for example, intuitions about "knows" over

intuitions about "justified"); (ii) avoiding areas of language upon which a lot of philosophical baggage has already accreted (perhaps a reason to prefer intuitions about "ignorant of p" over "knows that p"?); (iii) deploying substantively detailed settings for one's linguistic scenarios; (iv) dictionaries; and perhaps more interestingly, (v) attending to usage of the term in specialist areas, which for his topic of excuses he would include law and psychology ("with which I include such studies as anthropology and animal behaviour" (14)).

What Austin is not including there, under the rubric of psychology, are methods like those of some experimental philosophers, doing empirical work *on ordinary usage*, but rather seeing what sorts of distinctions and nuances the psychologists have devised in response to their own explanatory needs on the topics of responsibility and action. But it does seem to me that it would be reasonably within the spirit of the open-endedness of Austin's list of methodological ideas that it could include taking the temperature of "Clapham-omnibus types". I would also think it reasonable to add the use of Google as a low-cost way of getting comparisons of the frequency of different strings of words, while keeping in mind that it is a method with numerous potential pitfalls (see Kilgarrif (2007), though also Fletcher (2013) for more cautiously positive guidance).

Closely related to tin ear is the risk of *confirmation bias*, or what would be called experimenter bias in the context of scientific methodology, and perhaps could thus be called *intuiter bias* here. The earlier quote from Williamson is preceded by the following:

> After all, philosophers defending a given position against opponents have a powerful vested interest in persuading themselves that the intuitions that directly or indirectly favour it are stronger than they actually are. The stronger those intuitions, the more those who appeal to them gain, both psychologically and professionally. Given what is known of human psychology, it would be astonishing if such vested interests did not manifest themselves in at least some degree of wishful thinking, some tendency to overestimate the strength of intuitions that help one's cause and underestimate the strength of those that hinder it. If one tries to compensate for such bias effects, one may be led to undercompensate or overcompensate; the standpoint of consciousness gives one no very privileged access to whether one has succeeded, for bias does not work by purely conscious processes. Its effects are much easier to observe in others than in oneself. (p. 121)

As Williamson acknowledges, though, being aware of this source of error is perhaps of little help in correcting for it. This is not at all a jab at Williamson: his purpose in this passage was not at all to offer methodological guidance, but rather to point out that we are not always well-aware of what our own intuitions might be. But I believe his description of the psychology here, and its applicability to intuitions, is correct. Indeed, it may be worse than he suggests here, in that merely being aware of these sorts of biases can even exacerbate their effects (Babcock & Loewenstein (1997)).

4. How What We Don't Know Can Hurt Us: the Problem of Methodological Ignorance and Inferential Demand

The above sorts of errors are ones that have been uncovered by intuition-deploying philosophers themselves over the history of the practice. For at least some of them, we have

clear resources in our practices to anticipate, detect, and correct for them from within our armchair practices themselves. For others, this is not so obvious, such as with the "tin ear"/ intuiter's bias type of challenge.

Moreover, a new swath of sources of error have been made salient of late by both psychological research on judgment and intuition, as well as by the particular intra-philosophical version of such research known as "experimental philosophy". These are documented at greater length in chapter 22 so I will not rehearse them again here. Rather, let me just emphasize that that work is all still, even at its best, fairly preliminary, and (as is also discussed in that chapter) much debated. At the same time, the set of proposed possible effects, errors, and methodologically problematic sensitivities seems to be growing rapidly every year. This is still a very new area of research for philosophy. All in all, although a large number of plausible potential culprits have been identified and hypothesized about here, really what all this work so far suggests is, most simply, that *we just don't know very much about what really will or won't cause problems in particular applications of the appeal to intuitions.*

We are also in a state of fairly substantial ignorance as to what steps can be done to eliminate or at least mitigate these threats of error, where they may arise. One popular candidate, for example, has been that our training as philosophers may make a difference, and render the intuitive verdicts of the professional sufficiently less susceptible than those of the folk to all these kinds of errors (Ludwig, Hales, Williamson). But the short answer here is that *we really really just don't know* (Weinberg et al. 2010), although much recent work has suggested at least that philosophers are about as susceptible to some order effects as the folk are (Schwitzgebel & Cushman 2015), as well as to the effect of personality type (Schultz et al. 2011).[6] As I noted earlier, too much of the debate about intuitions has been framed in all-or-nothing terms, and one unfortunate side effect of that has been that defenders of the practice have not been on the whole too interested in the question of how to improve it.

The all-or-nothing, skeptic-versus-anti-skeptic frame of philosophizing about intuitions has also obscured a further question about what more we might require of our intuitive capacities than just reliability. If the only arguments that need to be responded to are skeptical ones denying any reliability to intuitions whatsoever, then it can seem that all that is required by way of a defense of intuitions in our methodology is that they be on balance reliable. I argue in my (2007) that there is further methodological norm on sources of evidence that I call *hopefulness*: we ought to trust a source of evidence only to the extent that we possess resources to detect and correct for errors that the source might be prey to. This is a requirement above basic reliability, such that even a highly-reliable-but-still-fallible source of evidence should not, I argued, be trusted if we lacked any good means of detecting and correcting for its errors as they might arise. I also argued that current philosophical practice with intuitions is pretty much devoid of such means, that is, in my technical term, *hopeless*. I will not deploy that argument here, but rather offer a different argument for the insufficiency of mere reliability as a methodological norm on sources of evidence,

[6] For recent reviews of this active research area, please see Buckwalter (forthcoming), Nado (2014) and Machery (2015).

for the extent of reliability that we ought to require of some source will depend on *what kinds of inferences* we may want to use based on the the evidence provided by that source.[7]

Goldman's classic statement of how to think about the standards of inferential justification is from his 1979: "A reasoning procedure cannot be expected to produce true belief if it is applied to false premises …. What we need for reasoning … then, is a notion of '*conditional reliability*'. A process is conditionally reliable when a sufficient proportion of its output-beliefs are true *given that its input-beliefs are true*" (emphases original; p. 13). Although Goldman is surely right that our norms for evaluating modes of inference should not be too dependent on the actual track record the procedures might have, given bad inputs, nonetheless his move here of stipulating that *all* of the input-beliefs are true obscures an important methodological distinction.

Consider the following two modes of inference. The rule TEN-CON takes 10 propositions as input, and infers to their conjunction, P1 & P2 & … & P10. TEN-CON is obviously as conditionally reliable as you could want: if all the inputs are true, then so is the output, always and necessarily. Compare that to TEN-DIS, which inputs 10 propositions, and outputs their disjunction, P1 v P2 v … v P10. This rule is also as conditionally reliable as one could want. Nonetheless, there is an important methodological difference between them: if even one of TEN-CON's inputs is false, then TEN-CON thereby cannot help but yield false outputs, but TEN-DIS will produce true outputs so long as at least one of the inputs is true. I will say that TEN-DIS is extremely *error-robust*, in that its utility as a mode of inference remains high even as the quality of its input set degrades. But TEN-CON is just as extremely *error-fragile*: its methodological value collapses quickly as any noise creeps into its premise set.

For rules of inference with small sets of premises, especially where each of the premises is highly likely to be true, there may be little to worry about regarding the difference between error-robust and error-fragile inference rules. With modus ponens, say, it might seem to be a distinction of insufficient utility to attend to at all. However, when one considers inferences where a great many data points will be considered, and when furthermore we have reason to suspect that a nontrivial number of those data may be in error, then this distinction becomes methodologically essential—under such conditions, a procedure which is error-fragile could be highly, even maximally, conditionally reliable, and still be such that we could have good reason to fear that its conclusions will be false much more often than not.

When pondering issues of inferential justification, epistemologists have not generally considered massively multi-premise inferences. But when we transition from epistemology to methodology, it becomes urgent to consider ways in which such large-scale inferences might work differently from the compact and small-scale inferences that we're more used to, and comfortable with, thinking about. At a minimum, I would suggest that something like the following methodological principle of error robustness (ER) holds:

> (ER) An inference procedure *I* can be legitimately trusted on a given premise set π only to the extent that *I*'s error-robustness is proportionate to the expected error rate in π.[8]

[7] See Gonnerman & Weinberg (forthcoming) for a fuller presentation of these arguments, in the context of an evaluation of Goldman's naturalistic and reliabilist account of the methodology of cases.

[8] If we were considering using some *I* across a number of different domains or contexts, where there may be very different sorts of premise sets π encountered, then we may want to add some appropriate relativizing structure to (ER). But I will not worry about that here.

That is, the more that we expect errors actually to crop up in π, the more error-robust we will require *I* to be, in order for us to trust *I*'s deliverances as to the consequences of π. The greater error-fragility of *I*, the greater requirement on any π that it be not *merely* on-balance reliable. And for highly error-fragile modes of inference, we may need to be equally adept at weeding out errors in their input streams. This principle follows quickly once we recognize that we do not just want our inferences to be reliable under conditions of ideally accurate premises, but rather we want them to be reliable under actually prevailing methodological conditions.

Moreover, the typical practices of philosophical inference that operate on the premises delivered by appeal to intuitions, appear to be highly error-fragile. At least in such highly intuition-driven subfields as epistemology, metaphysics, and the philosophy of mind, an operative norm seems to be something close to one good intuitive counterexample being enough to apply very heavy pressure to reject the theory being counterexampled. Here is Brian Weatherson in his (2003), contending that these norms hold broadly in philosophy; he makes these observations, I should note, as a prelude to his own case that we should move towards norms that would allow for more errors in our intuitive premises:

> In epistemology, particularly in the theory of knowledge, and in parts of metaphysics, particularly in the theory of causation, it is almost universally assumed that intuition trumps theory. Shope's *The Analysis of Knowledge* contains literally dozens of cases where an interesting account of knowledge was jettisoned because it clashed with intuition about a particular case. In the literature on knowledge and lotteries it is not as widely assumed that intuitions about cases are inevitably correct, but this still seems to be the working hypothesis. And recent work of causation by a variety of authors, with a wide variety of opinions, generally takes the same line: if a theory disagrees with intuition about a case, the theory is wrong. In this area exceptions to the rule are a little more frequent, particularly on the issues of whether causation is transitive and whether omissions can be causes, but in most cases the intuitions are taken to override the theories. Matters are quite different in ethics. It is certainly not a good thing for utilitarian theories that we very often feel that the action that maximizes utility is not the right thing to do. But the existence of such cases is rarely taken to be obviously and immediately fatal for utilitarian theories in the way that, say, Gettier cases are taken to be obviously and immediately fatal for theories of knowledge that proclaim those cases to be cases of knowledge. Either there is some important difference here between the anti-utilitarian cases and the Gettier cases, a difference that justifies our differing reactions, or someone is making a mistake. I claim that it is (usually) the epistemologists and the metaphysicians who are wrong. In more cases than we usually imagine, a good philosophical theory can teach us that our intuitions are mistaken. (2)[9]

While the nature of philosophical evidence has been a matter of much attention in recent years, nonetheless the nature of philosophical inference or theory-selection is more or less still not on our methodological radars. If the arguments in this section are correct, though, we really cannot properly do the former without at the same time doing a bit more of the latter.

[9] See also Nado (2015), and Weinberg (forthcoming).

5. On Intuitions and Armchairs

I have been contending here that we are operating in a current state of substantial ignorance about intuitions—or, speaking more carefully, about what (if anything) fills the intuition-role in our current practices, and what sorts of errors it is or is not susceptible to under what circumstances. But if it is true that philosophers' modes of inference using intuition are largely error-fragile, then it becomes methodologically pressing that we be relieved of this ignorance. This presents one crucial distinction between philosophical appeals to intuition as discussed here, and the ways in which intuitions of a different sort may be standardly used in the formal sciences: the latter have evolved with an enormous focus on not allowing any errors into its premises in the first place, in large part by a rather radical paring down of the sorts of intuitions that can be appealed to. That's part of what it was for mathematics and logic to become *formal* sciences in the first place. Such a move does not seem available for the rest of philosophy, though. Moving forward, intuition-deploying philosophy will also likely need to be scientific-psychology-deploying philosopy as well.

Now, we might very reasonably ask whether it is really any sort of *cost* to philosophy, if it turns out that we can continue using intuitions responsibly only if we pretty much forsake the armchair to do so. There may well be *practical* costs, in that it would obviously be much more cumbersome, difficult, and literally expensive to run empirical studies regarding how we come to make judgments about cases, than it is to just consider the cases ourselves and see what our minds present to us. The studies themselves would take substantial effort to run, but even more than that, acquiring the requisite training and expertise to run them in the first place would likely prove even more demanding. (I do not mean to suggest that the appeal to intuitions is effortless on the whole: devising the right sort of case to do the right sort of argumentative work can often be very challenging. But these costs of examining our intuitions about some case empirically would be imposed even beyond the initial costs of devising philosophically interesting cases to consider in the first place.) Let me offer three reasons to think that such practical costs may not be too great, after all. First, we would likely have a division of labor in which the great majority of intuition practitioners would not need to pursue this empirical work themselves, but would only need to be appropriately conversant with the relevant empirical literature as it developed. Second, and relatedly, although it would be very hard for philosophers with established careers to go back and learn how to do the science, for at least some philosophers nearer the start of their professional education, perhaps even from their undergraduate years, acquiring that skill set would be easier, even fun. Third, I do not expect that we would need to run studies for every single case for which philosophers might wish to appeal to an intuition. Rather, as we come to have a better sense of the error profile of these intuitive judgments, we will be able to take proper proactive steps and make good error-catching and error-avoiding amendments to current philosophical practice. And where a specific case becomes a matter of dispute, the profession will be able to call upon a stock of appropriately trained philosophers (and, I hasten to add, philosophically engaged psychologists as well, of which there are many excellent ones already in evidence).

So, perhaps the practical costs will not be too great, if we want to keep our intuitions at the cost of losing our armchair. But what about philosophically deeper *epistemic* costs?

I can only think of two principled reasons why we might want to require a philosophical project to steer clear of all experimental work. If that project was truly foundational, in the manner of Descartes's *Meditations*, then it might indeed be epistemically illicit to appeal to empirical results when the very trustworthiness of the empirical was being treated as not yet secured. But I do not think that that is a popular project to pursue just now.

The other reason one might have to eschew experimental premises altogether is if one is concerned that using *contingent* premises would preclude drawing inferences to *necessary* conclusions, along similar lines to how the necessitation rule in a modal logic must be restricted if, say, an actuality operator is being used. And, presumably, whatever is learned from the experimental philosophers' studies are going to be contingent. Now, I think such concerns may, depending on the project, be justified about our use of contingent premises in general. But those concerns should not apply to the kind of work I am advocating here, which is aimed at illuminating the boundaries of the trustworthiness of our capacity for intuition in the first place. For if our powers of modal intuition are trustworthy at all, then it is a fact about us that sometimes the contingent fact that we are intuiting that *necessarily, P* really is good evidence for the necessity of P. If modal skepticism is false, then our minds must offer a contingent mirror of the realms of necessity. And it is exactly the structure—and flaws—of that mental mirror that such experimental philosophy work would be aiming to illuminate. We thus need not fear any modal enervation from using experimental results about our intuitive capacities in such a way.[10]

I should add that one may have good reasons not to rely on any *particular* piece of putative experimental evidence, if one has doubts about the methods being used there. None of this section is meant as a pitch for any extant work of experimental philosophy, including my own. I am considering here, rather, the prospects of a *future* set of philosophical practices that is simultaneously intuition-based and experiment-guided. Let me also emphasize that though I have been addressing primarily the mainstream armchair practice of appeal to intuitions, many of these concerns also carry over wholesale to work in "positive program" experimental philosophy (Alexander, Mallon, and Weinberg 2010). The advantage that such experimental philosophers may have over their armchair brethren is not that the intuitions of the folk as gathered via survey are more trustworthy, either on the whole or seriatim, compared to those of the self-consulting intuition-monger. Rather, the advantage would come from availing themselves of the wider array of tools available to them for both illuminating the epistemic profile of those intuitions, and for taking steps to inoculate them from any sources of error they discover. For example, it is pretty much impossible to come to an understanding of order effects from consulting one's own intuitions, but experimental methods can first discern where and to what extent intuitive verdicts are sensitive to order, and then undertake to gather those intuitions using a set of survey instruments using appropriately shuffled orders (as indeed is standard practice in the social sciences, when one is concerned about order effects in one's study).

To conclude: we do not have any good reason not to pursue, by empirical means, a rich error profile for our intuitive practices, including how to detect and manage whatever sorts of errors we may be at real risk for. And we do have some good reasons to engage in such pursuit, in that our current general ignorance about the workings of whatever-it-is

[10] See my 2013 for a more filled-out version of this argument.

that may serve the intuition evidential role does not sit easily next to the demands that our inferential practices place on that role. In matters methodological, ignorance is rarely bliss.

Acknowledgments

Many thanks to Joshua Alexander, Jonathan Jenkins Ichikawa, and Ben Jarvis for guidance on an earlier draft.

Bibliography

Alexander, J., Mallon, R., and Weinberg, J., 2010, "Accentuate the negative," *Review of Philosophy and Psychology*, *1*, 297–314.

Austin, J., 1956, "A plea for excuses: The Presidential Address," *Proceedings of the Aristotelian Society,* 1–30.

Babcock, L., and Loewenstein, G., 1997, "Explaining bargaining impasse: The role of self-serving biases," *The Journal of Economic Perspectives*, 109–26.

Bealer, G., 1996, "On the possibility of philosophical knowledge.," *Noûs*, 1–34.

Bealer, G., 1998, "Intuition and the Autonomy of Philosophy," in DePaul and Ramsey 1998, 201–39.

Bealer, G., 2002, "Modal Epistemology and the Rationalist Renaissance," in T. Gendler and J. Hawthorne (eds.), *Conceivability and Possibility*. Oxford: Oxford University Press, 71–125.

Bealer, G., 2004, "The origins of modal error," *Dialectica*, *58*, 11–42.

BonJour, L., 1980, "Externalist theories of empirical knowledge," *Midwest Studies in Philosophy*, *5*(1), 53–74.

BonJour, L. 1998, *In Defense of Pure Reason*, Cambridge: Cambridge University Press.

Buckwalter, W., forthcoming, "Intuition fail: Philosophical activity and the limits of expertise," to appear in *Philosophy and Phenomenological Research*.

Buckwalter, W. and Stich, S., 2014, "Gender and Philosophical Intuition", in Nichols and Knobe 2015, 307–16.

Cappelen, H., 2012, *Philosophy without Intuitions*, Oxford: Oxford University Press.

Chalmers, D., 2014, "Intuitions in philosophy: A minimal defense," *Philosophical Studies*, *171*, 535–44.

Cohen, S., 1988, "How to be a fallibilist," *Philosophical Perspectives*, 91–123.

Cummins, R., 1998, "Reflections on Reflective Equilibrium," in DePaul and Ramsey 1998, 113–27.

DePaul, M. and W. Ramsey (eds.), 1998, *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*, Lanham, MD: Rowman and Littlefield.

Davis, W., 2007, "Knowledge claims and context: Loose use," *Philosophical Studies*, *132*, 395–438.

Deutsch, M., 2010, "Intuitions, counter-examples, and experimental philosophy," *Review of Philosophy and Psychology*, *1*, 447–60.

Deutsch, M., 2015, *The Myth of the Intuitive*, Cambridge, MA: MIT Press.

Devitt, M., 2006, "Intuitions in linguistics," *The British Journal for the Philosophy of Science*, *57*(3), 481–513.

Fischer, E. and Collins, J. (eds), 2015, *Experimental Philosophy, Rationalism, and Naturalism: Rethinking Philosophical Method*, London: Routledge.
Fletcher, W., 2013, "Corpus Analysis of the World Wide Web", in C. Chapelle (ed.), *The Encylcopedia of Applied Linguistics*. London: Wiley-Blackwell, 339–47.
Fodor, J., 1964, "On knowing what we would say," *The Philosophical Review*, 198–212.
Fodor, J., 1997, "Special sciences: Still autonomous after all these years," *Noûs*, *31*, 149–63.
Fodor, J. 1998, *Concepts: Where Cognitive Science Went Wrong*, Oxford: Clarendon Press.
Gettier, E., 1963, "Is justified true belief knowledge?" *Analysis*, 121–3.
Goldman, A., 1979, "What is Justified Belief?" in G. Pappas (ed.), *Justification and Knowledge.* Boston: D. Reidel, 1–25.
Goldman, A., 2007, "Philosophical intuitions: Their target, their source, and their epistemic status," *Grazer Philosophische Studien*, *74*(1), 1–26.
Grice, H. P., 1975, "Logic and Conversation" in P. Cole and J. Morgan (eds.) *Syntax and Semantics, Vol. 3, Speech Acts*, New York: Academic Press, 41–58.
Hales, S., 2012, "The faculty of intuition," *Analytic Philosophy*, *53*, 180–207.
Horgan, T. and Henderson, H., 2001, "The a priori isn't all that it is cracked up to be, but it is something," *Philosophical Topics*, 219–50.
Ichikawa, J., 2010, "Explaining away intuitions," *Studia Philosophica Estonica*, *2*, 94–116.
Ichikawa, J. and Jarvis, B., 2013, *The Rules of Thought*, Oxford: Oxford University Press.
Jackman, H., 2005, "Intuitions and semantic theory," *Metaphilosophy*, *36*(3), 363–80.
Jackson, F., 1998, *From Metaphysics to Ethics: A Defence of Conceptual Analysis*, Oxford: Oxford University Press.
Jenkins, C., 2008, *Grounding Concepts: An Empirical Basis for Arithmetical Knowledge: An Empirical Basis for Arithmetical Knowledge*, Oxford: Oxford University Press.
Johnston, M., 1987, "Human beings," *The Journal of Philosophy*, 59–83.
Kilgarriff, A., 2007, "Googleology is bad science," *Computational linguistics*, *33*, 147–151.
Kitcher, P., 1978, "On Appealing to the Extraordinary," *Metaphilosophy*, *9*, 99–107.
Kornblith, H., 1998, "The Role of Intuition in Philosophical Inquiry: An Account with No Unnatural Ingredients," in DePaul and Ramsey 1998, 129–141.
Kripke, S., 1980, *Naming and Necessity*, Cambridge, MA: Harvard University Press.
Lewis, D., 1996, "Elusive knowledge," *Australasian journal of Philosophy*, *74*(4), 549–67.
Ludwig, K., 2007, "The epistemology of thought experiments: First person versus third person approaches," *Midwest Studies in Philosophy*, *31*, 128–59.
Machery, E., 2015, "The Illusion of Expertise", in Fischer and Collins 2015, 188–203.
McKenna, M. 2014. "Resisting the manipulation argument: A hard-liner takes it on the chin," *Philosophy and Phenomenological Research*, 89, 467–84.
Malmgren, A. S., 2011, "Rationalism and the content of intuitive judgements," *Mind*, *120*, 263–327.
Nado, J., 2014, "Philosophical Expertise," *Philosophy Compass*, *9*, 631–41.
Nado, J., 2015, "Intuition, Philosophical Theorizing, and the Threat of Skepticism," in Fischer and Collins 2015, 204–21.
Nagel, J., 2012, "Intuitions and experiments: A defense of the case method in epistemology," *Philosophy and Phenomenological Research*, *85*(3), 495–527.
Nichols, S., and Knobe, J. 2014, *Experimental Philosophy: Vol. 2*, Oxford: Oxford University Press.
Pust, J., 2000, *Intuitions as Evidence*, New York: Garland/Routledge.
Putnam, H., 1990, *Realism with a Human Face*, Cambridge, MA: Harvard University Press.

Saul, J. 2007, *Simple Sentences, Substitution, and Intuition*, Oxford: Oxford University Press.
Schulz, E., Cokely, E., and Feltz, A., 2011, "Persistent bias in expert judgments about free will and moral responsibility: A test of the expertise defense," *Consciousness and Cognition*, *20*, 1722–31.
Schwitzgebel, E., and Cushman, F., 2015, "Philosophers' biased judgments persist despite training, expertise and reflection," *Cognition*, *141*, 127–37.
Sopher, D., 1977, "Ordinary language," *Annals of the Association of American Geographers*, 625–6.
Sosa, E., 2007, "Experimental philosophy and philosophical intuition," *Philosophical Studies*, 132: 99–107.
Sosa, E., 2009, "A defense of the use of intuitions in philosophy," in D. Murphy and M. Bishop (eds.), *Stich and His Critics*, Malden, MA: Wiley-Blackwell, 101–22.
Stich, S., 1988, "Reflective equilibrium, analytic epistemology and the problem of cognitive diversity," *Synthese*, *74*: 391–413.
Symanski, R., 1976, "The manipulation of ordinary language," *Annals of the Association of American Geographers*, 605–14.
Talbot, B., 2013, "Reforming intuition pumps: When are the old ways the best?" *Philosophical studies*, *165*, 315–34.
Weatherson, B., 2003, "What good are counterexamples?" *Philosophical Studies*, *115*(1), 1–31.
Weatherson, B., 2014, "Centrality and marginalisation," *Philosophical Studies*, *171*(3), 517–33.
Weinberg, J., 2007, "How to challenge intuitions empirically without risking skepticism," *Midwest Studies in Philosophy*, *31*, 318–43.
Weinberg, J., 2008, "Configuring the Cognitive Imagination", in K. Stock and K. Thomson-Jones (eds.), *New Waves in Aesthetics*. London: Palgrave Macmillan, 203–23.
Weinberg, J., 2013, "Experimentalist Rationalism, or Why It's OK if the A Priori Is Only 99.44 Percent Empirically Pure," in A. Casullo and J. Thurow (eds.), *The A Priori in Philosophy*, Oxford: Oxford University Press, 92–109.
Weinberg, J., 2014, "Cappelen between Rock and a Hard Place", *Philosophical Studies*, *171*, 545–53.
Weinberg, J., forthcoming, "Experimental Philosophy, Noisy Intuitions, and Messy Inferences", in J. Nado (ed.), *Advances in Experimental Philosophy and Philosophical Methodology*, New York: Bloomsbury.
Weinberg, J., Crowley, S., Gonnerman, C., Vandewalker, I., and Swain, S., 2012, "Intuition & calibration," *Essays in Philosophy*, *13*(1), 15.
Weinberg, J., Gonnerman, C., Buckner, C., and Alexander, J., 2010, "Are philosophers expert intuiters?" *Philosophical Psychology*, *23*, 331–55.
Williamson, T., 2004, "Philosophical 'Intuitions' and Scepticism about Judgement," *Dialectica*, 109–53.
Williamson, T., 2007, *The Philosophy of Philosophy*, Oxford: Blackwell.
Yablo, S., 1993, "Is Conceivability a Guide to Possibility?" *Philosophy and Phenomenological Research*, 1–42.
Yablo, S., 2006, "No Fool's Cold: Notes on Illusions of Possibility", in M. Garcia-Carpintero and J. Macia (eds.), *Two-Dimensional Semantics*, Oxford: Oxford University Press, 327–46.

CHAPTER 17

PHILOSOPHICAL PROGRESS

GARY GUTTING

1. Introduction

I agree with Iris Murdoch that "philosophy makes no progress"[1] in the sense that most people say this, and will try to explain why. But I will also try to explain just what is admirable—and in important senses, progressive—about philosophy's achievement.

When we ask whether philosophy progresses, we almost always have in mind a comparison with science or mathematics. Science is said to progress in the sense of gathering more and more data, of answering more and more questions, of solving more and more problems, of accumulating more and more truths, of producing an increasingly accurate picture of the world. (The first and last senses don't fit so well with mathematics.) The first three senses are pretty unproblematic, whereas the last two (which I do not intend to get into) raise flags, at least for those with worries about what truth is and whether it can "picture" the world.

Whatever the details of the story about science, it does progress in important ways and this progress is largely due to its empirical method—its ability to test claims against experience. It may be that philosophy too can test claims against experience; we will have to look at this possibility. But more often the thought is that philosophy is more like mathematics: a non-empirical discipline that achieves whatever it does simply by thinking.

I begin (section 2) by exploring this idea, reflecting on the Cartesian project—whether or not it is what Descartes himself had in mind—of beginning with a skeptically unassailable first truth and from there progressively building up a system of philosophical truths. I then present (section 3) a less problematic but similar project associated with contemporary analytic philosophy, noting, however, that it too fails to yield progress in answering the fundamental questions of philosophy. Section 4 examines the idea that philosophy might

[1] Iris Murdoch, *The Sovereignty of Good*, Routlege and Kegan Paul, 1970, 1. The full text reads: "It is sometimes said, either irritably or with a certain satisfaction, that philosophy makes no progress. It is certainly true, and I think this is an abiding and not regrettable characteristic of the discipline, that philosophy has in a sense to keep trying to return to the beginning: a thing which is not all that easy to do."

nonetheless progress in something like the manner of empirical science, never answering its fundamental questions but generating important intermediary results. Next, I suggest (section 5.1) that we give up the assumption—implicit in sections 2.1–4.1—that we need philosophy to ground our pre-philosophical convictions about fundamental questions ("philosophical foundationalism") and go on (section 6.1) to propose an alternative view of philosophy as providing rigorous theoretical formulations of general pictures. Working from this new model, I discuss in turn several issues highly relevant to the question of progress in philosophy: disagreement among philosophers (section 7.1), the role of intuitions in philosophy (section 8.1), philosophical knowledge (section 9.1), and the interaction of science and philosophy (section 10.1). Finally, in section 11.1, I present my conclusions about philosophical progress.

2. The Cartesian Project

Can we come to know things just by thinking? We do in mathematics. By just thinking about whole numbers we can see that each one has a successor, and by starting from simple truths like that we can logically derive other truths. So could philosophy be a progressive discipline that comes to know things in some way similar to mathematics?

This is an attractive idea, which has, in various ways, enchanted many philosophers from Plato and Aristotle, through the medievals, to the early modern rationalists. There is, however, a standard formulation of the view (often attributed to Descartes) that has soured many philosophers on it. On this formulation, we need philosophy to refute skepticism, even about beliefs that are most basic and obvious. The project of philosophy should begin by questioning everything, but nonetheless somehow arrive at one or more first truths that withstand all doubt, from which philosophers (or more specialized investigators in particular sciences) could derive all other truths. But repeated attempts to carry out this Cartesian project have shown that hardly any beliefs (except maybe logical tautologies and perhaps a limited version of the cogito) can withstand radical skepticism, and that we have little hope of proving anything of interest from this meager base.

At the same time, philosophers, beginning at least with Hume, have convincingly argued that we do not need philosophical justification to accept the basic beliefs needed to defeat skepticism. This leaves open in principle the quasi-mathematical project of starting from uncontroversial truths and trying to argue (by uncontroversial principles of deductive and inductive logic) for further conclusions.[2]

Here many will object that there are serious problems with any candidates for "uncontroversial" truths. So, for example, the idea that there are "self-evident" truths, clear to anyone who understands their meaning, or truths about the immediate givens of consciousness have both been subject to serious criticism.

[2] I say "quasi-mathematical" because the starting truths include some that are about the external world and not derived simply from thinking. But the arguments for further truths proceed entirely from thinking, not, like science, from observation and experimentation.

But such criticisms either attack particular accounts of uncontroversial truths or else merely remind us that no truth is absolutely certain (and so could in the future be given up). We still all accept a large and varied set of obvious truths. These include deliverances of sense experience (the existence of a material world and many truths about the things that make it up), basic logical and mathematical truths, the existence of consciousness and many truths about its contents, and many non-empirical truths (about—depending on which account we accept—concepts, meanings, or language).

But aren't there philosophers who have seriously questioned most if not all of these "obvious" truths? There are idealists who deny the existence of matter, eliminative materialists who deny the existence of consciousness, nominalists who reject all manner of concepts and meanings (linguistic or non-linguistic). But such philosophical accounts do not challenge the obvious truths of everyday life on their own terms.

Idealists, for example, do not fail to navigate around the things that ordinary people call "material objects", nor do eliminative materialists not believe (in the ordinary sense) the views they defend. In all such cases, philosophers are not denying the common-sense truth (which would be insane), but are rejecting what they regard as faulty philosophical interpretations of these truths. Berkeley did not deny that rocks exist: he merely rejected the move from their existence to a metaphysics of rocks as substances. The Churchlands do not deny that they think; they merely refuse to regard thinking as proof (or evidence) that there is something immaterial. The obvious truths of common sense about things and thoughts are independent of controversial philosophical accounts of them. There is no sane idealist who would deny that G. E. Moore had hands. There is no sane materialist who will deny that Moore thought he had hands.

3. The Analytic Project

Once we sweep away confusions due to irrelevant philosophical disputes, there remains a large body of truths that are entirely obvious. In principle, it may be possible to argue from these truths to non-obvious conclusions about controversial philosophical questions. This, I suggest, has been the implicit project of analytic philosophy over the last century. It was, I believe, not the project of most of the great philosophers of the past. They tended to think that they had access to special philosophical intuitions about truths beyond the obvious truths of everyday life and developed philosophical systems based on these intuitions. I'll have more to say about such intuitions. My question now is how successful this project of analytic philosophy has been.

It's easy to dance around the question by distinguishing various senses in which philosophy very likely progresses: by refuting certain arguments or theories, by developing new concepts and questions, by deepening our understanding of difficult issues. This is useful in its place, and I will return to such ideas. But everyone knows that the central question is whether philosophy progresses in the sense of definitively answering the great questions that are its primary subject matter: Are there any realities (God, the soul, Platonic Forms) beyond the ken of science? Is there any hope for personal existence after death? What are

the basic principles of morality? Are our actions free in any morally significant sense? What, if anything, is the ultimate point of human existence?

These are the sort of things philosophers since Plato have most wanted to know, and it's what most people outside of philosophy think that it's trying to discover. It's also entirely clear that philosophers have not definitively (or even tentatively) answered any of these questions. Different philosophers have their own preferred answers, and some even strongly believe that they have excellent reasons for them. But, in stark contrast to what progress means in science, philosophers as a group have not been able to agree on answers to the big questions. People in general cannot appeal to philosophy for reliable answers to these questions.

4. Philosophy as an Empirical Science?

But is science actually so different from philosophy? Perhaps we just have different standards. It seems that science has, no more than philosophy, provided answers to its deepest questions. What are the ultimate constituents of matter? How does life originate from the non-organic? What is the final theory that will provide a unified account of the microcosm and the macrocosm (that is, combine quantum theory and general relativity)? Why do the fundamental constants have the values they do?

What science presents as settled results are accounts that work at certain levels but fail if we push further. Is it any different in philosophy? We do not have a fully adequate definition of knowledge, but the justified-true-belief account gives correct results for almost any case that we are likely to encounter in ordinary life. Moreover, we have a general understanding of what goes wrong in Gettier-style cases: a justified true belief falls short of knowledge because the belief is justified as a matter of sheer luck. What we lack is a final account of precisely how to understand what's involved in "sheer luck". How different is this from the scientific status of the gas-law, PV = nRT, which often works as a good approximation, but breaks down when further variables cannot be ignored? We have ways of taking account of these variables, but no single simple formula catches the variation of P with V and T for all cases.

Similarly, in ethics utilitarianism offers a good account of many questions about how to behave, and has been elaborately developed to take account of various objections (e.g. the appeal to rule-utilitarianism). The same is true of Kantian deontology. What we lack is a single unified account that would somehow combine both into a single all-encompassing theory—of the sort Derek Parfit tries to develop in *On What Matters*. How different is this from the search of physicists, from Einstein to string theorists, for a unified field theory?

Finally, in the human sciences, particularly psychology, some of the deepest questions—for example, about the nature of freedom or of consciousness—seem to converge with basic philosophical questions; and, although both philosophy and psychology have contributed to discussions of these questions, neither (nor both together) have been able to find decisive answers.

Much more analysis of many more examples would be required to make a serious case that philosophical and scientific progress are significantly similar. But even if such a case

can be made, it remains true that the less than ultimate scientific results have a clear and enduring value for the community of non-scientists. It is much less clear that the corresponding philosophical results have a comparable value. Even a partial or approximate scientific understanding often has a great deal of practical value for controlling our environment. But what does philosophy achieve in, say, clarifying the nature of God, freedom, or the soul, if it does not tell us whether in fact God exists, we are morally responsible for our actions, or we will survive death? We shall see, however, that there is a fruitful way of thinking about philosophy—once we give up the idea that its primary function is to establish fundamental truths—that makes more sense of the analogy with scientific progress.

But before developing that viewpoint, we should first discuss the much stronger suggestion that philosophy can employ at least a quasi-scientific methodology. Here I have in mind the currently popular idea that philosophical theories are justified as the best explanations of the data in a given domain. The idea is that, like scientists, philosophers have bodies of data available—from ordinary experience and even scientific results—that are relevant to problems such as the nature of consciousness, free will, the existence of God, the principles of justice. Why not, then, construct philosophical theories that try to explain these data; that is, general accounts from which we can derive (deductively or probabilistically) the data? As in science, the theory that best explains the data (that is, allows us to derive all or most of the data) should be accepted. If more than one theory does the job, we can, like scientists, appeal to other factors such as simplicity and coherence.

This approach has been particularly important in discussions of ontology. For example, should we accept scientific realism, the view that the (unobservable) theoretical entities posited by science exist and causally produce the phenomena we observe? Or should we treat such posits as mere "fictions" or "instruments for calculation" that have no actual existence? On one line of argument, assuming that the theoretical entities exist explains the fact that scientific theories positing such entities are so successful in predicting the outcome of experiments or other observations. So, for example, if heat is actually the motion of molecules, we can readily understand how the kinetic theory is able to predict the behavior of heated bodies. But if there are no molecules, the success of these predictions would seem to be a miracle. The realist philosophical theory is a better explanation than the anti-realist theory.

There is, however, a major difference between "inference to the best explanation" in philosophy and in science. In science, having found a theory that nicely fits all the data merely establishes the theory as a plausible hypothesis. The next step is to derive new observational claims from the theory and see if these novel predictions hold up. This is particularly important when more than one theory fits the initial data. But in any case, scientists realize that there's a good chance that a theory will not work when it's extended beyond the data it was constructed to explain. That's why fruitfulness (the continual generation of successful predictions) is a major criterion for theoretical success. In philosophy, however, there's usually little prospect of deriving new predictions from a theory, since we generally already know the relevant facts about what we want to explain.

What, for example, does the hypothesis of scientific realism predict beyond the success of scientific theories, which is precisely what it is posited to explain? This is a especially instructive example, since the scientific case for the existence of any particular theoretical entity is far stronger than the philosophical case for the existence of theoretical entities in

general. A physicist postulates electrons to explain certain known phenomena and then goes on to predict as yet unobserved phenomena that we should expect if the theory is true. It is the indefinitely repeated success of such predictions (fruitfulness) that establishes the truth of the theory. Once we have the scientific evidence for the existence of various theoretical entities, a philosophical case for the existence in general of all entities posited by highly successful theories is otiose. In fact, the best proof a philosopher can offer that theoretical entities exist is to point to the evidence that science has produced for their existence.

5. Against Philosophical Foundationalism

This brings us, then, to the thought that we need to reconceive the primary function of philosophy. Suppose the project of philosophy as an intellectual discipline—committed to developing a rationally based consensus about the big questions—disappeared or had never existed. What difference would that make? Would, for example, people not have beliefs about God, freedom, immortality, and morality? Of course not. Once humans have reached a certain level, they will develop religions, codes of behavior, political systems, etc. that come with beliefs about such matters. All they would lack would be the peculiar idea (not accepted by even most people in modern Western society) that such beliefs can or ought to be established by arguments of the sort philosophers try to develop.

A number of recent philosophers—for example, Wittgenstein, Rorty, and Plantinga—have developed strong cases for rejecting this foundationalism. But there is a more straightforward way to make the point. Philosophy has not justified our fundamental religious, ethical, or political beliefs. Given this and philosophical foundationalism, it follows that we should give up all these beliefs that guide us at the deepest level through our life in the world, the beliefs that define our moral identity. The absurdity of thinking that we should do this is sufficient reason for rejecting philosophical foundationalism.

To make the point concrete, think of the beliefs that informed the life of someone you greatly admire: perhaps Socrates, Jesus, Voltaire, Jefferson, or Gandhi. Can you really agree that such a person should have given up the beliefs that gave rise to that life because they weren't justified by philosophical arguments? There remains the interesting question, to which current discussions are contributing, of just why disagreement among philosophers should not require skepticism. But, for present purposes, we can rest with the obvious fact that skepticism is not required, and from it conclude that philosophical foundationalism is false.

But a world without foundational philosophy does not mean a world with no thinking about the fundamental beliefs of religion, ethics, and politics. At a minimum, people will be interested in better understanding their cherished beliefs, deriving their logical consequences, and—in an intellectually diverse society—answering challenges from those who reject their beliefs. Because of this, a world that has no need for a philosophical grounding of its beliefs may, nonetheless, profit immensely from contact with a tradition of philosophical thought.

An excellent example is medieval Europe, with its widespread faith in Christian revelation. The re-emergence of Aristotle's works provided an enormous resource for the basic

Anselmian project of "faith seeking understanding". (There was, of course, also a strong and earlier influence of Platonic philosophy.) Admittedly, Thomas Aquinas, for example, did accept Aristotle's first principles as self-evident truths and believed that important parts of the Christian belief-system (the existence and nature of God, natural-law morality) could be established by philosophical reasoning from these truths. But Aquinas insisted that such philosophical support was not necessary, just an optional though valuable supplement to faith, which itself provided both the natural and supernatural truths essential to Christianity.

Apart from this optional foundational role for philosophy, Aquinas found in Aristotle a variety of conceptual distinctions and modes of argumentation that were crucial tools in his development and defense of Christian faith. The distinction between substance and accident, for example, enabled him to explain how what seemed to be bread and wine were in fact the body and blood of Christ. Similarly, the Aristotelian practical syllogism was a fruitful means of developing more fully the moral content of the Ten Commandments. At the most general level, the Aristotelian picture of the world (with some revisions, for example, concerning Creation and the eternity of the world) provided an intellectually comfortable context for Christian thinking.

I suggest that, for all our differences from the medievals, we can fruitfully see philosophy playing a similar essential role in our thinking. In our pluralistic society, there is a considerable variety of basic convictions, especially about religion, science, ethics, politics, and art. If we take our own convictions seriously (and have the education and time), we are likely to have an interest in clarifying, developing, and defending them against criticism—in other words, an interest in the *intellectual maintenance* of our convictions. Most of the intellectual resources for this maintenance come from what philosophers have achieved over the last 2,500 years.

The advantages of philosophy are most readily apparent when professional philosophers interact with those not trained in the discipline, for example, undergraduates or even professionals in other disciplines. We philosophers may like to think that we have an advantage in such discussions because of our superior intelligence. In fact, however, it's because we know more philosophy than our interlocutors. Think of the difference it makes in discussions of, say, God's existence to distinguish between gratuitous evil and evil necessary for a greater good; in discussions of freedom to introduce the idea of compatibilism; in discussions of ethics to appeal to rule-utilitarianism; or in discussions of democratic political theory to have Rawls' distinction between comprehensive and overlapping consensus.

6. Pictures and Theories

To better understand what philosophy provides for our intellectual maintenance of prephilosophical convictions, I propose a distinction between *pictures* and *theories*. Our convictions are typically not very well articulated and can be best regarded as expressing *pictures*; that is, general schemes for thinking about some major aspect of the world. Some of the most important current pictures are naturalism, materialism, theism, libertarianism, determinism, utilitarianism, and deontologism. These are expressions of

many non-philosophers' basic convictions and the subjects of philosophical reflection. Specifically, one of the main projects of philosophical thinking is the developing of precise and detailed formulations of such pictures. Such formulations are *theories*.[3]

Philosophers may start from a particular picture because it expresses their own basic convictions, or they may be interested in calling the corresponding convictions into question, or the picture may have just attracted their intellectual curiosity. In any case, once we have an apparently plausible, sufficiently detailed theoretical formulation of a picture, the next task is to see if we can refute it. There are various techniques of refutation: showing that the theory is self-contradictory, showing that it is inconsistent with obvious truths, showing that it has no intelligible content (e.g. is based on a distinction that makes no sense). Once a theory is refuted, a philosopher can either try to formulate a better theory expressing the picture or develop a new picture. Formulating a new theory is a matter of revising the old one in light of its refutation. This means not only dropping the claims that have been shown false, but also replacing them with claims that both avoid the refutations and seem plausible in their own right. Here it is common to move through several successive theories, particularly though the dialectic of counterexamples. Kripke's critique of descriptivism is a classic example of the theoretical development of a picture for the sake of refutation. Rawls uses his method of reflective equilibrium to make a positive case for his picture of justice as fairness.

Once a picture has shown a capacity for inspiring viable theories, the development and continuing viability of these theories becomes the central locus of what I call the "persuasive elaboration" of the picture. Ideally, philosophers would be able to arrive at a theory that survived all criticism and could, moreover, be shown superior to all rival theories. Such a theory would stand as a prime example of philosophical knowledge.

But as we know, philosophers almost never achieve this sort of theoretical success. Theories either require successive modification in the light of ever new difficulties or, at least, cannot be shown to be decisively superior to rival theories. But successful theoretical development does show that a given picture is capable of generating progressively better theoretical formulations. Such success may demonstrate, for example, an ability to meet counterexamples and other objections, or an ability to extend the picture to new domains. Philosophers can and do agree about the present status of the theoretical development of a given picture (even though they may themselves hold different views about the ultimate acceptability of the picture). For example, after Kripke the picture of metaphysical necessity as an objective fact is far more viable than it was before; the theistic picture and the dualist (anti-materialist) picture are in much better shape after the work, respectively, of Plantinga and Chalmers; and Rawls significantly challenged the utilitarian picture.

[3] I'm adapting the distinction of pictures and theories from Kripke's comments in *Naming and Necessity*. The distinction is flexible and relative, since pictures and theories differ merely in degree of specificity. What functions as a theory in one context (e.g. Plato's "theory of Forms" in the *Phaedo*) may function as a picture in another (e.g. discussions of universals among contemporary analytic metaphysicians). The value of the distinction—like that of any distinction—lies not in its universal correctness but in the light it is able to shed on particular cases to which it can be applied. This and the next five paragraphs are largely taken from Chapter 10 of my *What Philosophers Know*, Cambridge University Press, 2009.

Results of this sort are obviously important for those holding the relevant convictions or beliefs, but they can also be significant for those who do not. If the picture associated with the convictions is shown to be internally consistent, to cohere with well-established bodies of knowledge, and to suggest interesting and fruitful responses to questions important even to those not sharing the convictions, then even non-believers may well need to take these convictions seriously. For example, a materialist reading Chalmers or a utilitarian reading Rawls may, though still unconvinced, become much less dismissive of these convictions.

Efforts at theoretical development of a picture can also lead to refutations of convictions or, conversely, defenses of convictions against purported refutations. The latter is well exemplified by Plantinga's free-will defense against the deductive argument from evil, a defense that makes a strong case for the logical compatibility of an all-powerful, all-good God and the existence of evil. Convictions can also be decisively refuted by, for example, showing them to be logically contradictory or inconsistent with established facts. So, for example, simple versions of mind–body dualism collapse when it's shown that they can make no sense of interactions between the mental and the physical. But convictions are typically formulated in terms of pictures that can take on a variety of specific theoretical formulations, so that refuting one such formulation will not decisively discredit the picture or the conviction. On the other hand, pictures and associated convictions that persistently fail to generate defensible theories will be rightly judged non-viable.

Elsewhere I've offered a fuller example of how philosophical results can be essential for intellectual maintenance of various religious (and anti-religious) pictures.[4] Here I want to connect my account of the role of philosophy to several major meta-issues: the significance of disagreement among philosophers, the role of intuition in philosophy, the existence of philosophical knowledge, and the relation of philosophy to empirical inquiry. A survey of these issues will provide a basis for my concluding discussion of philosophical progress.

7. Philosophical Disagreement

The disagreement of philosophers has always been the primary external challenge to the cognitive authority of philosophy: if they can't even agree among themselves, people say, why should we pay any attention to them? In the last few years, analytic philosophers have raised the question from within philosophy itself, asking whether I, an individual philosopher, should ascribe any special authority to my own conclusions about disputed philosophical questions, given that other philosophers (my epistemic peers, as we say), as knowledgeable, intelligent, and careful as I, disagree with me. Some philosophers—Peter van Inwagen, for example—see no good response to this sort of worry.[5]

Van Inwagen focuses on the example of his disagreement with David Lewis about the truth of incompatibilism, van Inwagen thinking he has a decisive argument against it,

[4] *What Philosophers Know*, 231–40.

[5] Peter van Inwagen, "We're Right. They're Wrong," in Richard Feldman and Ted A. Warfield (ed.), *Disagreement*, Oxford University Press, 2010, 1–28.

Lewis demurring. Van Inwagen rejects the skeptical conclusion that he (and Lewis) should respond to their disagreement by withholding judgment on the issue. He is convinced that, somehow, they both have an epistemic right to accept their own (mutually exclusive) views.

But van Inwagen also sees an unanswerable argument against this view. To be rational, he needs to pay attention only to the evidence for and against a claim, since it is only evidence that we have any reason to think tracks the truth. But he and Lewis are working from the same body of evidence regarding the truth of compatibilism. Van Inwagen can rationally hold incompatibilism only if the evidence points to its truth. But if the evidence points to the truth of incompatibilism, then it cannot be rational for Lewis to deny it. So, since van Inwagen holds that Lewis rationally rejects incompatibilism, he must conclude that the evidence does not point to the truth of incompatibilism. But then it follows that van Inwagen is not rational in accepting incompatibilism. In short, given that they are epistemic peers working from the same evidence, van Inwagen and Lewis cannot rationally hold opposing views. They should both withdraw to skepticism regarding the issue. Van Inwagen, however, simply reports his unwillingness to make this withdrawal.

There are various distinctions and strategies, short of van Inwagen's voluntarist refusal, that try to avoid the skeptical conclusion. But, as we should expect, there is no agreement as to whether any of these work. Suppose they don't, and suppose the conclusion is that philosophers are not, in general, entitled to their views on disputed questions. Why should this be a problem?

It is a problem only if we accept what I have called philosophical foundationalism; that is, if we think that, without a philosophical justification, we are not entitled to hold our fundamental religious, ethical, political, and aesthetic beliefs. But, as we saw in section 5, we should reject philosophical foundationalism.

This rejection allows us to loosen the connection between philosophizing and searching for rationally justified answers to the big questions. Philosophers may have such answers as an ultimate goal. But they, like everyone else, need not (and should not) wait for such an uncertain achievement before they endorse life-directing beliefs. They may devote themselves, for example, to the intellectual maintenance rather than justification of such beliefs. But they may also explore the internal logic of beliefs for which they have little sympathy, or try to give a more rigorous formulation to a picture that merely fascinates them intellectually. (Here we would better speak of intellectual *engagement* than of intellectual *maintenance*.) Philosophers who take these latter approaches need have no ultimate commitment to the philosophical conclusions they wind up defending. They can present their philosophical conclusions as simply their opinions on disputed matters, not necessarily preferable to those of peers with contrary views.

Of course, nothing I have said eliminates in principle the possibility that philosophy could some day discover decisively argued solutions to the great questions that are its ultimate focus, thereby providing a solid rational foundation for comprehensive philosophical consensus. To reject philosophical foundationalism is merely to deny that a rational foundation is necessary for us to be justified in our fundamental beliefs. It leaves us to pursue such a foundation as an ultimate ideal.

8. Philosophical Intuitions

Philosophical foundationalism requires us to derive answers to the fundamental questions through rigorous argument. This means that we need basic premises sufficient to support such substantial inferences. Such premises cannot, under pain of infinite regress, be themselves known by argument, so they must be somehow known directly. Let's call premises that require no argument "intuitions". There would be no problem if we could produce the needed arguments from utterly obvious premises such as "I exist", "x = x", and "there are material objects in motion". Then the philosophical "intuitions" would be truly uncontroversial truths that anyone who understands them readily accepts. Some claim that even the most seemingly obvious truths require philosophical defenses against radical skeptical challenge. But there is far more reason to doubt this claim than there is to doubt the obvious truths (e.g. of mathematics and sense perception) the claim calls into question.

The problem is, rather, that the intuitions required to carry out the foundationalist project are not obvious—even to those who understand them—and so themselves become subject to philosophical debate. This doesn't affect individual philosophers who are merely developing their own basic beliefs or exploring the basic beliefs of others, but it stops in its tracks the foundationalist program of providing generally compelling answers to fundamental questions.

Nonetheless, individual philosophers' developments of their own intuitions can have significance for the philosophical community as a whole. What an individual philosopher takes as a compelling intuition and uses as a premise in the derivation of further conclusions, the wider community can regard as a hypothesis that the philosopher is developing to show, for example, its systematic fruitfulness for making sense of an entire domain of inquiry. What the individual sees as deductive proof from a firmly accepted premise, the community can profitably view as hypothetico-deductive support for the premise. This is the process of persuasive elaboration that I mentioned earlier. As we've already noted, lacking predictions of unexpected results, this process will not amount to a decisive proof of the hypothesis. But it can make the community as a whole appreciate the relevance and plausibility of a view it has not been taking seriously. A premiere example is Kripke's development in *Naming and Necessity* of his intuitions that names are rigid designators and that, contrary to Quine, things have essential properties. Other examples are Alvin Goldman's reliabilist epistemology and David Lewis's contextualism.[6]

Apart from the foundationalist project, there is no need to develop an account of philosophical intuitions beyond the straightforward view that they are premises that are judged to be obviously true, along with the distinction between those that are obvious to just about everyone and those that only some see as obvious. Given that we cannot derive answers to fundamental questions from premises everyone takes as obvious, the foundational project requires us to try to make a case for special "philosophical" intuitions, which those specially trained in the discipline can see to be obvious. Husserl's phenomenology and the logical analyses of the twentieth-century positivists are examples of this enterprise.

[6] I discuss these three example at in some detail in *What Philosophers Know*.

Unfortunately, all such projects collapsed when no consensus about the special intuitions emerged among philosophers.

There is, of course, room for philosophical discussions about the nature and origin of our experiences of what we take to be "obvious" (in need of no reason supporting its truth). There is also room for challenges to the truth of what "everyone" thinks is obvious. But there is no avoiding the fact that rational thought must start from what, at the moment, we regard as obviously true. Even arguments challenging the "obvious" need to start from premises that we regard as obvious.

It is, therefore, a mistake to think that, in general, we are entitled to appeal to intuition only if we first show that intuition is reliable. A proof of reliability is needed for claims that the intuitions of a particular individual or of a special philosophical methodology are reliable. But intuitions, in the basic sense of truths that everyone accepts, need no justification. If they did, there would be no place to begin philosophical (or any) discussion.

9. Philosophical Knowledge

What do philosophers know? That is, what truths can philosophy as a discipline legitimately put forward to the public as established? Richard Rorty lampooned the very idea of such knowledge:

> Imagine . . . that a few years from now you open your copy of the *New York Times* and read that philosophers, in convention assembled, have unanimously agreed that values are objective, science rational, truth a matter of correspondence to reality, and so on . . . By way of making amends for the intellectual confusion which the philosophical profession has recently caused, philosophers have adopted a short, crisp set of standards of rationality and morality. Next year the convention is expected to adopt a report of the committee charged with formulating a standard of aesthetic taste.
>
> Surely the public reaction to this would not be "Saved!" but rather "Who on earth do these philosophers think they *are*?" It is one of the best things about the life we Western intellectuals lead that this *would* be our reaction.[7]

But what's absurd about the situation Rorty imagines is the idea that anyone would take seriously a mere consensus of philosophers—the mere fact that this particular group happened to agree about something. That, of course, would have no more standing than a consensus of flat-earthers or of astrologers. Suppose, however, that the philosophers' consensus emerged from detailed no-holds-barred discussions, in the course of which almost everyone, even those previously opposed, was convinced of, say, the compatibility of freedom and determinism. Suppose, further, that, subsequently, anyone who seriously challenged the philosophers' views and engaged them in discussion came to conclude that the philosophers were right. Then we would have as much grounds for thinking there is a body of philosophical knowledge as we do for thinking there is a body of, say, chemical knowledge.[8]

[7] Richard Rorty, *Objectivity, Relativism, and Truth*, Cambridge University Press, 1991, 43–4.

[8] It might seem that the case of a science such as chemistry is different because its conclusions are supported by publically available empirical evidence, whereas philosophy's are not. But even in the case of an empirical science, the empirical evidence supports the discipline's cognitive authority only

There is in fact such a body of philosophical knowledge. Some of it is historical knowledge of the current status of theoretical work on competing pictures. For example, as we've noted, mind-body dualism, metaphysical essentialism, and the logical compatibility of God and evil are far more defensible since the work of Chalmers, Kripke, and Plantinga. The model for evaluating philosophical work is similar to Imre Lakatos's "methodology of scientific research programs", with the strength of a general picture depending on the ability of the picture to generate successively more and more adequate theoretical formulations. Success in this regard, is the measure of a picture's "progress".

Even more important, though less recognized, is the vast number of conceptual distinctions that philosophers have discovered and refined over the centuries. Of course, everyone makes distinctions all the time: it's virtually identical with thinking. Philosophical distinctions are those that are central for discussing the fundamental questions to which philosophy seeks answers. We tend to ignore these distinctions, first, just because they are so common as to be unremarkable and, second, because, if we do think of them, they seem to be merely means to what philosophy is really after: answers to our fundamental questions.

But once we give up philosophical foundationalism and recognize that a great deal of philosophical thinking is devoted to intellectual maintenance of (or engagement with) prephilosophical convictions, it becomes apparent that distinctions are the life-blood of this sort of thinking. My religion teaches me that my soul is immortal. But is there a distinction between the soul and the mind or between a soul and a person? And what sort of relation does the soul have to the body? Is it part/whole or substance/property or agent/instrument? What could it mean to say that the soul is immortal? Does it mean that the soul exists at all times or that it exists outside of time (is eternal)? All these questions are put in terms of distinctions that have a long and complex philosophical history, and it is impossible to discuss them in any depth without drawing on this history—or else taking enormous time and pains to develop crude approximations to what's already available in that history.

It will be said that philosophers disagree about distinctions just as much as they disagree about the fundamental questions the distinctions are deployed to discuss. There are famous cases, such as fact/value and analytic/synthetic, where philosophers disagree about whether there is any distinction at all, and many others—material/immaterial, knowledge/justified true belief, free/compelled, for example—where they argue about just how to draw the distinction.

Such disagreements make a difference when matters are pushed far enough, particularly if we are after decisive philosophical answers to fundamental questions. If, for example, we want philosophical knowledge about whether we are free in a sense that allows for moral responsibility, then we will need a final account of the distinction between a free action and one that is compelled. But short of such a final account, there are various ways of drawing the distinction that can help clarify, defend, or criticize particular formulations of a prephilosophical conviction that we are free in a sense that allows for moral responsibility. We don't, for example, need a full explication of the freedom/compulsion distinction to recognize the mistake of neuroscientists who think they have disproved free will by finding neural events that correlate with our choices or who fail to distinguish between functional

because people agree that the evidence is as the scientists say it is and that the evidence is relevant to whether the scientific conclusions should be accepted. If there were a similar consensus about the nonempirical evidence for philosophical claims, philosophy would have a similar cognitive authority.

and phenomenal consciousness. Similarly, discussions of the problem of evil profit from the distinction between necessary and gratuitous evil, even without a final account of the nature of necessity.

Here we can return to the analogy with approximate laws in natural science. Even if the distinction between fact and value has no ultimate significance because (for example) on the deepest level to be desirable is just to be desired in a certain way, it remains true that, above this level, there is a meaningful difference between my desiring a drink and the drink being desirable. Similarly, even if we cannot ultimately show an essential difference between analytic and synthetic truths, there is still in many cases a highly relevant distinction between, say, "2+3=5" and "It is raining". Distinctions, like laws, can be excellent guides in many situations, even if they are not strictly true. Philosophy knows many highly useful distinctions and, when they reach their limits, it is likely to have suggestions for further refining them.

There is, then, a body of positive, first-order philosophical knowledge. It has emerged, since Plato, from philosophers' continuing efforts to answer the fundamental questions, to achieve the wisdom that defines their enterprise. The knowledge consists of hundreds, perhaps thousands, of conceptual distinctions. Many of them are nascent in ordinary language and thought, but philosophers have developed them with detail, refinement, and rigor that go far beyond the ordinary. This development is essential to the goal of answering the fundamental questions of philosophy.[9] But, as I have emphasized, they are also essential for everyone—philosopher or not—who wants to clarify, extend, defend, or critically refine or revise their own pre-philosophical answers to these questions. These distinctions constitute the established body of knowledge that gives philosophy its own authority as a cognitive discipline. They are the truths non-philosophers can rely on us to provide.

10. The Interaction of Philosophy and Science

Scientific research is another context in which philosophical knowledge can be crucial. This is particularly true in the social sciences and psychology, where empirical inquiry is often directed to questions that turn on conceptual issues. I've already mentioned scientific results purporting to show that there is no free will. There are, for example, experiments that identity neural events that precede by up to three-quarters of a second what agents perceive phenomenologically as a choice and allow us to predict the choice with up to 80% accuracy.[10] But, apart from methodological problems with such experiments, concluding from them that we are not free would require an understanding of what it means for a choice to be free (in the sense appropriately tied to moral responsibility). There are no

[9] For a good discussion, see Robert Sokolowski, "The Method of Philosophy: Making Distinctions", *The Review of Metaphysics* 51 (1998), 515–32.

[10] For a survey of such results, see Kerri Smith, "Neuroscience vs. Philosophy: Taking Aim at Free Will," *Nature* 2011 (477), 23–35.

experiments that could determine this: it requires philosophical knowledge of the distinction between free and unfree in a morally relevant sense.

This is not to say that neuroscience is not relevant to the question of free will. The experimental study of the brain can give us specific information about how brain events affect our choices. The philosophical task of understanding freedom may be greatly aided by focusing on the real neurological situation, instead of mere abstract possibilities. Even if philosophers will never arrive at a full understanding of what, in all possible circumstances, it means for a choice to be free, they may, working with brain scientists, learn enough to tell when, if at all, the choices we make in ordinary circumstances are free.

We may be a long way from any results as significant as what this simple example suggests might be possible. But there is already a great deal of intellectual cash being generated by the joint work of scientists and philosophers in the more philosophical areas of physics (the interpretation of quantum theory and relativity), biology (the definition of species, the understanding of evolution), and cognitive science (consciousness studies). We should expect more such cooperation as philosophers come to see the value of working from real examples, and scientists—especially in the human sciences—see the extent to which they need philosophical distinctions to understand their results.

11. Philosophical Progress

As I said at the outset, when we ask about philosophical progress, we typically have in mind a comparison with science. We have seen that philosophy, like science, is a successful cognitive enterprise: it has generated a substantial and significant body of knowledge. The question, then, is whether this knowledge, like science, shows a progressive character.

There's an obvious minimal sense in which any successful cognitive enterprise makes progress: there is, over time, more knowledge than before. So, in philosophy, we continually formulate new distinctions, refine old ones, and improve the theoretical formulations of various general pictures. But the scientific model of progress is much stronger. Scientists begin with a set of questions, discover satisfactory answers to them, the answers raise new questions, and the process continues. The result is not just more knowledge but the systematic accumulation of knowledge, leading to (at least as a limit) a complete account of a certain domain.

There are traditions in certain periods of the history of philosophy that exhibit something like this stronger sort of development. Platonism, Thomism, British empiricism, Kantianism, logical positivism, and phenomenology have, during certain periods, developed in a quasi-scientific cumulative way. But at some point the development stops and the tradition splits into warring camps or simply peters out. Revivals may occur, but with the same results. In any case, a tradition remains just one strand in the history of its period, never commanding general assent among philosophers.

The question of progress takes a particularly disputatious turn with regard to the history of philosophy. The progress of science leaves its history behind. It goes without saying that no working scientists are concerned with outdated work such as geocentrism or the phlogiston theory of combustion. But even work accepted as permanently significant (Newton's

mechanics and optics, Maxwell's electrodynamics) is formulated in up-to-date ways that require no reference to the original texts of their discoverers. Apart from *pietas* or an idiosyncratic interest in history for its own sake, a contemporary scientist needs no history beyond the names of the great dead that are used to designate constants, laws, or theories.

Although there are analytic philosophers who reject the history of philosophy as irrelevant to their contemporary enterprise, courses in history remain an essential part of the both undergraduate and graduate education, and, more important, many outstanding analytic philosophers see historical work as an integral part of their search for philosophical knowledge. In fact, there has been a contemporary renaissance in historical work led by scholars who combine analytic acuity and historical sensibility.

Robert Pasnau nicely connects this emphasis on history to the question of philosophical progress. There is a dilemma, he says: "Either philosophy has progressed over the centuries or it has not. If it has not, then what good is philosophy? If it has, then what good is its history?"[11] I suggest that the answer is that philosophy has not progressed in the manner of the sciences, where progress means the essential irrelevance of a discipline's history. But, as we have seen, philosophy provides a body of knowledge that is good (indeed, essential) for our intellectual engagement with pre-philosophical convictions.

Why, given philosophy's cognitive success, doesn't it move beyond a methodological reliance on its history? Why, in other words, must it keep rethinking the ideas of the past? The answer to this question should help us see why philosophy has not progressed in the manner of science.

The questions philosophy is ultimately trying to answer are the fundamental questions of human existence. Human beings require answers to at least some of these questions to lead meaningful lives and cannot wait for the deliberations of philosophy to provide them with full rational rigor. That is why, as I've emphasized, philosophy takes place in a world filled with pre-philosophical convictions. From the standpoint of a rational inquiry into fundamental questions, such convictions are "prejudices"—not silly beliefs that need to be abandoned but "pre-judgments" about the topics of philosophy that need critical assessment.

Contrary to the hopes of the Cartesian project, there is no way of simply freeing ourselves from all the convictions of our society and reaching the totally disinterested "view from nowhere". But we can and should challenge our local convictions with those of other societies, and for this purpose the study of past philosophers—at least our intellectual peers and working in contexts of quite different convictions—is a rich resource. It is also valuable to compare our Western modes of philosophical thinking with similar intellectual efforts in other civilizations and also to confront philosophical thinking itself with alternative modes of engaging with basic convictions, such as myth, literature, and history.

This need for continuing engagement with the history of philosophy signals our difficulty in attaining sufficient intellectual "distance" to provide rationally compelling answers to the fundamental questions that define our discipline. Some philosophers, particularly in the "continental" tradition hold that attaining such distance is impossible. In

[11] Robert Pasnau, "Philosophical Beauty". Paper read at Harvard-ANU Conference of Philosophical Progress and Intellectual Culture (available at <http://spot.colorado.edu/~pasnau/research>, accessed September 20, 2015).

any case, the philosophical enterprise is particularly challenging because it seeks rigorously rational answers to questions that concern the fundamental meaning of human life. To ask about God, freedom, and immortality is not just to inquire as to whether a certain state of affairs obtains: it is also to ask about the ultimate significance of our existence. We might, then, say that philosophical questions are hermeneutic, not merely factual.

Science has been able to progress in a strong, cumulative sense because it has focused on a set of factual questions that can be answered without getting into questions about what they ultimately mean for human existence. These answers often provide means of controlling the world for our purposes. But they are neutral as to what these purposes (derived from what we take to be the meaning of our lives) ought to be. Scientists sometimes do try to move from their factual results to conclusions about the fundamental hermeneutic issues (claiming to have shown that God does—or does not—exist, to have discovered what happiness is, etc.); but then they become embroiled in all the ambivalences and uncertainties of philosophy. The distinctive progress of science depends on its sticking to factual, not hermeneutic, questions.

In conclusion, then, does philosophy progress? No, not in the primary sense of achieving a cumulative series of results whereby, on the model of science, it provides increasingly complete and adequate answers to its fundamental questions. In a weaker sense, however, the steady increase in philosophical knowledge of distinctions and of the strengths and weaknesses of various pictures and their theoretical formulations is an important form of progress.

Finally, our discussion suggests two other senses in which philosophy is at least associated with human progress. First, philosophy often contributes to the interpretation and clarification of scientific results, and to that extent shares in the progress of science. Second, as an essential resource for our intellectual engagement with our pre-philosophical convictions, it is essentially involved in whatever progress (e.g. moral or political) that we may regard human civilization to have achieved over the centuries.

References

Gutting, Gary, *What Philosophers Know*, Cambridge: Cambridge University Press, 2009.

Murdoch, Iris, *The Sovereignty of Good*, London: Routledge and Kegan Paul, 1970.

Pasnau, Robert, "Philosophical Beauty". Paper read at Harvard-ANU Conference of Philosophical Progress and Intellectual Culture (available at <http://spot.colorado.edu/~pasnau/research>, accessed September 19, 2015).

Rorty, Richard, *Objectivity, Relativism, and Truth*, Cambridge: Cambridge University Press, 1991.

Smith, Kerri, "Neuroscience vs. Philosophy: Taking Aim at Free Will," *Nature* 2011 (477), 23–35.

Sokolowski, Robert, "The Method of Philosophy: Making Distinctions," *The Review of Metaphysics* 51 (1998), 515–32.

van Inwagen, Peter, "We're Right. They're Wrong," in Richard Feldman and Ted A. Warfield (eds.), *Disagreement*, Oxford: Oxford University Press, 2010, 1–28.

CHAPTER 18

CONCEIVABILITY AND POSSIBILITY

CHRISTOPHER S. HILL

1. Introduction

It is widely recognized, both in philosophy and everyday life, that some propositions are not just true but *necessarily* true.[1] They are propositions that could not have been false, that would have been true no matter how the world had been arranged. To give some examples, it is widely acknowledged that the laws of logic are necessary, and that the same is true of principles that are constitutive of concepts, such as the proposition that green is a color, and the proposition that a fortnight is a period of fourteen days. There is also wide agreement that the laws of mathematics are necessary.

In addition to recognizing a category of necessary propositions, most of us also recognize that some propositions are *possible*. This second category includes all true propositions, including all of the propositions that are necessarily true, but it also includes many propositions that are false, such as "I am not typing right now" and "I do not exist." Even though I am in fact typing right now, I might not have been typing, and even though I do exist, I might not have existed. Possibility can be defined in terms of necessity: to say that a proposition is possible comes to the same thing as saying that it isn't necessary that the proposition is false.

Necessity and possibility are known as *modal* features of propositions. There are two other types of modal status that should be noted. A proposition *P* is said to be *contingent* if it is possible for *P* to be true and also possible for *P* to be false, and *P* is said to be *impossible* if the denial of *P* is necessary.

There are many important philosophical questions about the nature of modality, but there are also significant questions about knowledge of modality. This chapter is concerned

[1] I am indebted to my student David Black for helpful conversations. I have also been helped by comments on an earlier version by two referees for Oxford University Press.

with the latter questions. How do we determine whether a proposition is necessary or contingent? What procedures do we use for recognizing possibility?

Traditionally, philosophers have held that—to a very large extent, anyway—we achieve modal knowledge by relying on the faculty of *conceiving*, that is, on our ability to conceive of objects and situations. We know that a person might not have existed by conceiving of a situation in which the parents of the person remain childless, and we know that it is necessary that 3 is the immediate successor of 2 by trying to conceive of a number between 2 and 3 and realizing that it cannot be done. Philosophers have also traditionally held that *the imagination* plays a significant role in the acquisition of modal knowledge. In some authors, this view is closely related to the view that modal knowledge derives from conceiving, for they use "imagine" so broadly that states of conceiving are included in its sphere of application. But other authors distinguish between conceiving and imagining, maintaining that the former necessarily involves the use of concepts, and is therefore best understood as akin to thought and judgment, while the latter always involves imagery of some kind. According to these authors, there are at least two distinct ways of gaining modal knowledge. Unfortunately, I will not be able to discuss this view here. There is room only to discuss one form of modal knowledge, and conceiving is the natural choice. Conceiving is a more comprehensive source of modal knowledge than the imagination, for we are able to represent more possibilities by using concepts than by using images.

I will be concerned to characterize conceiving, to determine whether it is sufficient to explain our modal beliefs, or is merely a component of a larger package, to consider whether its deliverances are reliable, and to assess the prospects of employing it to settle an important metaphysical issue, the mind–body problem. These topics will be the primary focus of the chapter, but to do justice to them, it will be necessary to attend to some metaphysical questions. Hence, I will also be concerned to some extent with modality itself, considering its different forms, how those forms might be analyzed, and why they are important to us.

2. Conceiving

To conceive of a moose in a suit purchased from Brooks Brothers is to form a complex representation involving the concepts of a moose, Brooks Brothers, a suit, the relation of being purchased from, and the relation of wearing. The representation will also have a logical structure. Thus, for example, the concept of a suit and the concept of Brooks Brothers will be bound respectively to the first and second argument positions of the concept of being purchased from. This example illustrates how conceiving works in general: to conceive of something is to form a representation of it that has concepts as its building blocks and the structural relations of logical grammar as its cement.

As this characterization implies, it is possible to conceive of anything that can be represented by binding together concepts in a way that conforms to the rules of logical grammar. Thus, for example, it is possible to conceive of a sample of water that is not composed of H_2O, a rectangle that is a perfect ellipse, and a barber who shaves all and only those men who do not shave themselves.

It follows that simple, undisciplined conceiving is not a reliable test for possibility. On the present account of conceiving, it is possible to conceive of *anything*, including logical contradictions. If we are to have a reliable test for possibility, we must rely instead on what I will call *constrained conceiving*—conceiving that is compatible with the laws of logic, the principles that are constitutive of concepts, and any other propositions that are assumed to be necessary in the relevant context. The need to impose constraints on simple conceiving has not always been acknowledged in the history of philosophy, but there are many anticipations of the point. Thus, for example, Descartes maintained that we might easily be misled if we rely on conceivings that are confused or insufficiently articulated. This observation is correct, but it does not go far enough. In addition to checking for clarity and distinctness, it's at least necessary to consider conceptual representations in relation to the laws of logic, and in relation to the principles that spell out the contents of the relevant concepts, with a view to determining whether the representations are compatible with these laws. Another way to put this is to say that in order to show that *X* is possible, where *X* is an object or a state of affairs, one must construct a model of *X* in thought, and then determine whether the model satisfies the laws of logic and the principles that are constitutive of the relevant concepts. Of course, given that our resources of time, memory, and attention are quite limited, any argument that a model satisfies the relevant principles will be fragmentary and therefore of heuristic value at best. But conceiving must involve a certain amount of checking on constraints if it is to be trustworthy even to a modest degree.

3. A Priori Modalities

To see how these ideas work in practice, let us focus on a conception of modality that has arguably been the primary focus of a number of philosophers, including Descartes, Hume, Kant, and the logical positivists. On this conception, the concept of a necessary truth is coextensive with the concept of a proposition that can be known to be true a priori. Propositions are regarded as possibly true if they are compatible with the necessary truths.

Advocates of this conception have generally allowed that we can gain knowledge of possibility by conceiving, but this makes sense only insofar as conceiving is held to be subject to a constraint. Suppose that one conceives of a situation by forming a conceptual representation *R*. Then it must be true that *R* is compatible with all of the categories of a priori propositions that are constitutive of necessity. In the interests of simplicity, I will assume here that there are exactly two such categories—the laws of logic and the principles that are constitutive of concepts (hereafter the *laws of concepts*). Thus, to establish that a situation is possible, one must construct a representation of the situation in thought and determine that it is compatible with the a priori propositions of these two types.

It would be interesting to develop this appraoch at greater length, but I wish to focus here on an epistemological problem that arises already in connection with the present barebones summary. The summary offers an account of how we come to know possibility, but it provides no explanation of how necessity is recognized. It is assumed at the outset that all and only a priori propositions are necessary. But how can that assumption be justified?

How can we come to know that it is true? Not by constrained conceiving. Since the assumption is built into the definition of the relevant form of constrained conceiving, any attempt to use that form of conceiving to establish the assumption would be circular. Nor can we hope to establish it by an argument based on informal characterizations of a priority and necessity. To say that a proposition is a priori is to say that it is possible to *know* it without relying on information about *empirical* facts. On the other hand, to say that a proposition is necessary is to say that it is *true* under *all* possible arrangements of facts, whether empirical or non-empirical. As Kripke emphasized, there are no immediate logical connections between these ideas (Kripke 1980, Kripke 2011). We can conceive of a proposition's being known independently of empirical facts without conceiving of it as true under all possible arrangements of facts, and we can also conceive of a proposition's being true under all such arrangements without conceiving of it as knowable independently of experience (by any being who is remotely akin to us). To say that the two concepts are coextensive is to make a substantive claim that cannot be established by a simple argument. Moreover, given the differences between the two concepts, it is very hard to see what a complex proof of coextensiveness would look like.

Faced with this difficulty, one might be tempted to explain our knowledge of necessity by saying that it arises from an explicit definition that reduces the concept of necessity to non-modal concepts. Thus, one might propose to *define* necessity by saying that a proposition is necessary just in case it is a law of logic or a law of concepts. This particular proposal is not very attractive, because, like all purely disjunctive definitions, it provides no basis for explaining why its disjuncts should be grouped together under a common label. But one might more reasonably propose to define necessity by saying that a proposition is necessary just in case it is known a priori. This definition would significantly reduce the task of explaining modal knowledge, for it is comparatively easy to see how one can be justified in believing a definition. After all, a definitional belief merely records a stipulation that is imposed by the will of the believer. Unfortunately, reflection shows that this approach is unsatisfactory. Any account of necessity must vindicate the intuition that necessity is a matter of truth relative to all possible configurations of facts, and no reductive definition in terms of non-modal concepts can achieve that result. Thus, for example, the definition in terms of a priority implies that there is no more to the idea of necessity than being knowable independently of experience. But this seems wrong. When one asserts that a proposition that is known to be a priori is necessary, one isn't just making the uninteresting claim that it's a priori. One is making a substantive claim about the world. Or so it seems.

Here is another reason for doubting that necessity can be explained reductively in terms of a priority. We know that all of the laws of logic are necessary. But do we know that all of the laws of logic are *actually* known a priori? No. No creature who is subject to the resource constraints that govern actual human cognition can be said to have an exhaustive knowledge of logic. To be sure, we can conceive of an ideal being who is capable of knowing all of the laws of logic, but we cannot be sure that such a being exists in the actual world. The most we can be sure of is that it is *possible* for such a being to exist. Hence, if we are to explain necessity in terms of a priority, we must say something to the effect that a proposition is necessary just in case it is *possible* for there to be a being who knows the proposition a priori. It is clear that this definition is not reductive. Thus, it makes explicit use of a modal notion, and moreover, a modal notion that belongs to the same family as the modal

notion that is being defined. (The notion of necessity that is being defined is the traditional Cartesian notion, and it therefore satisfies the condition that it is possible to know all necessary truths a priori. The notion of possibility that figures in the definiens must satisfy the corresponding condition—that is, the condition that all knowledge of possibility is a priori.)

Can we say that we acquire knowledge of necessity by some special cognitive faculty that might be called *modal intuition*? No. Since we have no independent reason to believe in modal intuition, such a move would be question-begging, and anyway, it would provide us with no explanatory insight. It would be like saying that we are apprised of the truth of mathematics by a special faculty of mathematical intuition. That move has had its defenders in the history of philosophy, including Plato and Gödel, but most philosophers have held it in low esteem.

At this point, one might be tempted to dismiss the classical package consisting of a priori modality and the associated form of constrained conceiving as confused. Perhaps it should be jettisoned in favor of some alternative package. But this proposal is also unacceptable. For the significance of the problem is quite general. Let *N* be any collection of truths that are thought to be necessary. We can test for possibility by conceiving of situations that satisfy the members of *N*, but how will we show that the members of *N* really do have the modal status that has been claimed for them? We will not be able to argue that all efforts to conceive of counterexamples to the members of *N* result in failure, for as we have seen, it is possible to conceive of anything at all. Nor will we be able to establish that the members of N are necessary by showing that all attempts to construct counterexamples via *constrained* conceiving come to grief. The relevant form of constrained conceiving will presuppose that the members of *N* are necessary, and this will preclude using that form to give a noncircular justification of the necessity of those propositions.

To summarize, if we are not to wind up as skeptics about modal knowledge, we must recognize that constrained conceiving is not the only way of determining the modal status of propositions.[2] But we must also recognize that it would not help to try to explain modality via reductive definitions based on non-modal concepts. Such definitions can explain how we gain knowledge of necessity, but they inevitably fail to do justice to its distinctive metaphysical nature, which has to do with truth under all possible configurations of facts.

[2] As a referee has pointed out, it might be useful to restate the argument for this point in somewhat different terms.

To determine whether a situation is possible, we typically start by *conceiving* of the situation, where this means that we combine concepts to form a representation of it. But this is not enough, because it is possible to form a conceptual representation of virtually *any* situation, including one that has features that are straightforwardly contradictory. (For example, it is possible to form a conceptual representation of this coin's being both round and square—as is illustrated by this very sentence.) To complete the exercise, we must see whether our conceptual representation will pass through a *filter* that consists of the necessary propositions. This means that we must have some way of determining which propositions are necessary. Can we make such determinations by some sort of conceivability test? At first sight, it might seem that the answer should be "Yes." Thus, we might hope to determine whether a proposition *P* is necessary by trying to conceive of a situation in which *not-P* is true, with failure in this enterprise being taken as grounds for attributing necessity to *P*. On reflection, however, it becomes clear that this procedure is badly flawed. As we've just observed, it's possible to conceive of *anything*. Hence, even if *P* is necessary, an attempt to conceive of a situation in which *not-P* is true

Nor can we deal with the problem by positing a new cognitive faculty. That approach has certain advantages, but they are the advantages of theft over honest toil. Nor can we deal with it by putting the traditional conception of modality aside in favor of an alternative conception.

With a view to solving this problem, or at least making some progress toward its solution, let us return to the intuitive characterization of necessity, which says that a proposition is necessary if it is true relative to all possible configurations of facts—or in other words, true in all possible worlds. According to this characterization, modal knowledge is either implicitly or explicitly knowledge of possible worlds, and modal reasoning is reasoning that leads from some claims about possible worlds to other such claims. Now in order to describe a possible world it is necessary to use concepts, and by the same token, it is necessary to presuppose any propositions that may be constitutive of the concepts one deploys. After all, there can be little point in deploying a concept if one is not prepared to grant the truth of the principles that determine its content and guide its use. It follows that in describing a world, one must presuppose that the laws of logic hold in the world, for we will need to use logical concepts in any such description, and the laws of logic are constitutive of our logical concepts. Further, one must presuppose that all of the principles that are constitutive of non-logical concepts hold in the world. This is because, for every concept *C*, it is necessary to use *C* in giving a full description of a world, even if nothing in the world falls under *C*. Suppose, for example, that a world contains no spherical objects. It will be necessary to mention this fact in describing the world, and this means that the concept of a sphere must figure in the description. Now the concept of a sphere is governed by a number of principles, including the principle that a sphere is a three-dimensional figure, and the principle that every point on the surface of a sphere is equidistant from its center. It would make no sense to describe an object as a sphere unless one was prepared to see it as conforming to these principles. Accordingly, they will be presupposed by any description of the world.

This line of thought can be summarized by saying that the laws of logic and the laws of concepts must be *partially constitutive of the concept of necessity*, because they will perforce be presupposed in any attempt to characterize the truth-makers of modal claims. It follows from this that *if* we are ever epistemically entitled to believe *any* propositions that are concerned with necessity, *then* we are entitled to assume that the laws of logic and the laws of concepts are necessary. Of course, this conditional falls short of implying the categorical proposition that we *are* epistemically entitled to attribute necessity to the laws. Whether that holds depends on whether we are epistemically entitled to believe *any* propositions involving the concept of necessity. I will not be able to deal with this larger issue here. My present ambition is more modest. I am assuming that knowledge of

would meet with success. Can we proceed by first forming a conceptual representation of a situation in which *not-P* is true, and then attempting to pass this representation through a filter of necessary propositions? No. That would beg the question at issue, which is how we can determine which propositions should go into the filter.

I conclude that no test based principally on conceivability can be adequate to determine the modal status of all propositions. There must be some way of identifying necessary propositions that is largely or entirely independent of conceivability.

necessity is *possible*. My concern is only to explain *how* it is possible. That is, my goal is to identify ostensible sources of modal knowledge, and to assess possible impediments to trusting them.

If we wish to develop these views systematically, we will begin by saying that the following principle is the first component of an *implicit definition* of the concept of necessity:

(a) If a proposition is necessary, then it is true relative to all possible configurations of facts.

Next, we will offer the following principles as additional clauses of the implicit definition:

(b) If *P* is a law of logic, then *P* is necessarily true.
(c) If *P* is a law of concepts, then *P* is necessarily true.

Other clauses of the implicit definition will include the axioms of a modal logic, perhaps S5. (There will be some redundancy here, since the first component of the implicit definition implies the standard modal axiom to the effect that a proposition is true in the actual world if it is necessary, and the second component is equivalent to another standard modal principle.) Our third step will be to point out that all of the components of the implicit definition are epistemically available to anyone who possesses the relevant concept of necessity. If *any* claims involving that concept count as objects of knowledge, they will have that status as well, because a reflective grasp of them is a precondition of any legitimate use of the concept. And fourth, we will observe that despite being characterized by independent axioms, the concept of necessity is unified by the fact that the laws of logic and the laws of concepts enjoy a common nature. They would be presupposed in any attempt to specify the possible configurations of facts that serve as truth makers for modal claims.

Unlike the explicit definitions of necessity that we considered earlier, this implicit definition makes no attempt to reduce the concept of necessity to non-modal concepts. On the contrary, it makes explicit use of a modal concept in characterizing necessity. (Cf. principle (a).) In effect, on the present conception, the notions of necessity and possibility mutually constrain one another, due to their interactions in the clauses of the implicit definition.

It remains to be determined whether principles (b) and (c) are sufficient to specify the main categories of necessary propositions. What about the class of mathematical truths? I think we must draw a distinction in order to answer this question. On the one hand, there are propositions that are partially constitutive of mathematical concepts. Examples include the proposition that the successor function is one-one and the proposition that sets are individuated by their members. On the other hand, there are substantive claims about the existence of mathematical objects, such as the principle that every natural number has a successor and the axiom of choice. Members of the former class (which I will henceforth refer to as *constitutive mathematical laws*) can appropriately be grouped with the laws of logic and the laws of concepts, for they too determine the contents of their constituent concepts and provide guidance in their use. It makes sense to expand the implicit definition to include a clause that represents them as necessary. But it is not at all clear that members of

the latter class are on a par with the laws of logic and the laws of concepts. And by the same token, it is not at all clear that they should figure in an implicit definition of the concept of necessity.

In sum, while simple, undisciplined conceiving is not a reliable source of modal knowledge, it appears that constrained conceiving can be used with some confidence to establish that propositions are possible, where constrained conceiving consists in showing that a logically structured conceptual representation is compatible with the laws of logic, the laws of concepts, and certain of the laws of mathematics. To be sure, one cannot hope to establish conclusively that a representation is compatible with all of these laws, for there can be no complete proof procedure for establishing logical consistency. However, one can at least give heuristic arguments for compatibility claims of the relevant sort. We have also found that knowledge of necessity must come from some source other than constrained conceiving. Plausibly, it derives from certain axioms that provide a conceptual framework for describing the possible worlds that serve as truth-makers for modal claims. These axioms are clauses of an implicit definition of the concept of necessity. Since they have this status, one is entitled to believe them simply in virtue of possessing that concept, provided that one is entitled to hold *any* beliefs in which the concept plays a role. It is perhaps the best argument for this approach that most other ways of explaining our epistemic entitlement to beliefs about necessity quickly come to grief. (For an exception, see the Appendix. For more discussion of epistemic entitlements arising from implicit definitions, see Boghossian 1997.)

4. Metaphysical Modalities

In a dramatically sharp break with the traditional view, Saul Kripke and Hilary Putnam argued forcefully in the 1970s that necessity does not coincide with a priority. (Kripke 1980, Putnam 1975) There is a large class of necessary propositions that are not a priori. In defending this assertion, Kripke and Putnam relied primarily on appeals to intuition, presenting their readers with examples, and urging that their claims about the examples were vindicated by conceivability tests. It turned out that their readers shared these assessments, as did the subjects in a range of imaginative experiments conducted by the psychologists Frank Keil and Susan Gelman (Keil 1989, Gelman 2003).

Generalizing from the examples that Kripke and Putnam presented, we have the claim that the *true* members of the following categories are both necessary and a posteriori:

(i) Propositions consisting of proper names and the identity predicate
(ii) Propositions about the origins and ancestry of human beings
(iii) Propositions about the origins of artifacts
(iv) Propositions about the original material constitutions of physical objects
(v) Propositions attributing atomic numbers to atoms
(vi) Propositions attributing molecular structures to chemical kinds
(vii) Propositions attributing genetic structure to biological kinds

To illustrate, this view entails that the proposition that Mark Twain was identical with Samuel Clemens is necessary despite being a posteriori, and that the same is true of the proposition that Queen Elizabeth II came from a certain sperm and egg, the proposition that George VI was the father of Elizabeth II, the proposition that a certain rock was initially composed entirely of granite, and the proposition that water is composed of H_2O molecules.

In addition to arguing that there are necessary propositions that are not a priori, Kripke maintained that there are a priori propositions that are not necessary. His examples of the contingent a priori are controversial, but, inspired by his efforts, his readers soon identified other examples that have won a wider following. Thus, it is plausible that various indexical propositions fill the bill, including the proposition that I am here now, the proposition that I exist, the proposition that I am thinking, and the proposition that I am thinking that *P*, where *P* is a mathematical proposition, or any other proposition whose constituent concepts can be acquired independently of experience. I can know that I am here now simply in virtue of possessing the relevant indexical concepts, but it is entirely contingent that any particular person is in any particular place at any particular time.

On one interpretation of their work, Kripke and Putnam should be seen as criticizing the traditional view of modality—and in fact, as decisively refuting it. On this view, traditional philosophers like Descartes were concerned with the same type of necessity as Kripke and Putnam, but they had false views as to the propositions that enjoy that status, mistakenly holding that the propositions are limited to ones that can be known a priori. This interpretation enjoys a certain amount of plausibility, but there is also another interpretation that merits consideration. According to this second construal, Descartes and other traditional philosophers were concerned with a different form of necessity than Kripke and Putnam, a form that is possessed by a narrower range of propositions. On this interpretation, the Cartesian view is similar to the view put forward by Kripke and Putnam in that they both explain necessity in terms of truth relative to all possible configurations of facts; but this is merely a formal similarity, for the two views employ different conceptions of possibility, the Cartesian conception being much more inclusive than the Kripke–Putnam conception.

Now as I see it, there is ample motivation for both the Cartesian view of modality and the more recent Kripke–Putnam view. Thus, we are often interested in knowing what scenarios are mandated by the structure of our conceptual scheme, and what scenarios are allowed by that structure, and we are also often interested in knowing what scenarios are either mandated or allowed by the structure of our conceptual scheme in combination with assumptions about the essential natures of objects. Accordingly, I find it more plausible to suppose that we have two different conceptions of modality, and will presuppose this view in the sequel. In order to have names for the two types of possibility, I will refer to them respectively as *conceptual possibility* and *metaphysical possibility*. (I will refer to the two corresponding types of necessity as *conceptual necessity* and *metaphysical necessity*.)

How do we recognize metaphysical possibility? The answer is that we recognize it by constrained conceiving, but in this case, the constraints are more numerous. In addition to the laws of logic, the laws of concepts, and the constitutive mathematical laws, they include all propositions of types (i)–(vii), together with any other a posteriori necessities that may exist.

This proposal is quite different than the one that Kripke appears to favor. More specifically, as I read him, he appears to think that if we guard against certain possible confusions, we can achieve modal knowledge by making use of simple, undisciplined conceiving. Thus, for example, he appears to think that if we are careful not to confuse the proposition that Hesperus is identical with Phosphorous with a proposition that picks out Hesperus and Phosphorous by mentioning certain of their contingent properties, such as the proposition that the brightest heavenly body seen in the evening sky in the West is identical with the brightest heavenly body seen in the morning sky in the East, we will not be able to conceive of a situation in which the former proposition is false. Whether Kripke holds it or not, this view is clearly wrong. To conceive of a situation in which Hesperus is not identical with Phosphorus one need only construct a representation in which the concept of Hesperus and the concept of Phosphorus are bound respectively to the first and second argument positions of the concept of identity, and then prefix the concept of negation to the result. Like many philosophers, Kripke appears to have a view of conceiving that does not mesh with the psychological facts. To conceive is to form a conceptual representation. Period. As we observed earlier, given this understanding of conceiving it is possible to conceive of anything whatsoever, even including contradictions. It is only *disciplined* conceiving that provides an adequate test for possibility.

How do we recognize that an a posteriori proposition is a metaphysical necessity? According to Kripke, we do so by a process that is partly a priori and partly a posteriori. Focusing on the proposition that a certain table was not made of ice at its origin, he argues that the proposition is both necessary and a posteriori, and then offers the following account of how we recognize it as necessary (Kripke 2011):

> So we have to say that although we cannot know a priori whether this table was made of ice or not, given that it is not made of ice, it is *necessarily* not made of ice. In other words, if *P* is the statement that the lectern is not made of ice, one knows by a priori philosophical analysis, some conditional of the form "If *P*, then necessarily *P*." If the table is not made of ice, it is necessarily not made of ice. On the other hand, then, we know by empirical investigation that *P*, the antecedent of the conditional, is true—that this table is not made of ice. We can conclude by *modus ponens*:
>
> $$P \supset \Box P$$
> $$P$$
> ---
> $$\Box P$$

Generalizing from this case, we can conclude that for every proposition *P* that belongs to one of the foregoing seven categories (that is, categories (i)–(vii)), Kripke thinks that the conditional *If P, then it's necessary that P* can be known a priori, and that we come to know *It's necessary that P* by inferring this proposition from the conditional and the proposition that *P*, which we learn by normal a posteriori methods.

Kripke's discussion is illuminating, but it leaves us with the question, "How do we know the major premises of these arguments?" How do we know conditionals of the form *If P, then it's necessary that P*, where *P* is a proposition of one of the types (i)–(vii)? The answer I wish to recommend derives from one that was originally proposed by Alan Sidelle. (Sidelle 1989) According to Sidelle, we are entitled to accept propositions of the given sort

simply in virtue of possessing the concept of metaphysical necessity, for those propositions are partially constitutive of the concept. Building on this suggestion, I wish to propose that the concept of metaphysical necessity is implicitly defined by a rather heterogenous class of propositions that includes the laws of modal logic, appropriate versions of principles (a), (b), and (c) from section 3.1 (with the notion of conceptual necessity replaced by the notion of metaphysical necessity), the principle that a proposition is necessary if it is one of the constitutive laws of mathematics, and general principles that imply all conditionals of the sort that we have just been considering. Since these propositions are all components of an implicit definition of the concept, they are available to anyone who possesses it. Moreover, anyone who possesses it is epistemically entitled to believe them, provided that we are ever entitled to believe *any* propositions about metaphysical necessity.

This proposal has a certain prima facie appeal, but before it can be accepted, we must consider whether the clauses of the implicit definition have enough in common to yield a concept of necessity that is internally coherent. What, if anything, ties together the various a priori laws with propositions of types (i)–(vii)? On the face of it, there is a vast difference between, say, De Morgan's laws and the proposition that George VI was the father of Elizabeth II. What could unify a concept that owes its content to such different propositions?

In considering this question, we should recall that the various a priori laws are unified by the fact that they are all constitutive of the concepts that we must use in characterizing possible worlds. We presuppose them in describing the truth-makers of modal claims, so it is appropriate that they figure in an implicit definition of necessity. If a similar claim could be made about propositions of types (i)–(vii), there would be no obstacle to seeing the concept of necessity as unified. So we must ask whether there is a reason for thinking that propositions of the given types are presupposed in descriptions of possible worlds. And in fact, it seems that there is such a reason, for it seems that we presuppose propositions of types (i)–(vii) in assessing claims to the effect that actual objects and actual kinds exist in other possible worlds. Consider, for example, the claim that Elizabeth II exists in another world. Reflection suggests that we would not accept this claim unless we were convinced that there are individuals in other worlds who have exactly the same ancestors as Elizabeth. No matter how much a person in another world resembles Elizabeth in other respects, we will not allow that the person is identical with Elizabeth unless she has the same parents as Elizabeth, the same grandparents, and so on. In general, we will not say that actual people exist in other worlds unless the relevant propositions of type (ii) are satisfied. But this means that we presuppose true propositions of type (ii) in describing other worlds. In other words, those propositions are constitutive of the metaphysical modalities because we assume that they are true in characterizing the facts that serve as truth-makers for modal claims. Similar remarks apply to true propositions of type (i) and types (iii)–(vii).

This line of thought provides a rationale for supposing that all propositions of the form *If P, then it's necessary that P*, where *P* is a proposition of one of the types (i)–(vii), are constitutive of the concept of metaphysical necessity. But it is best not to think of them as separate constituents of an implicit definition of the concept, for on that construal of the definition, it would be quite unwieldy. It's better to suppose that the constituents of the definition are generalizations from which all such particular conditionals follow.

Here, then, is a theory that systematizes and explains the modal insights of Kripke and Putnam, while enabling an account of how we acquire modal knowledge—an account that is based on constrained conceiving and grasp of the clauses of an implicit definition.

There is, however, a question about knowledge of the metaphysical modalities that remains to be addressed. In contemporary philosophy the metaphysical modalities are held in high esteem, and knowledge of facts involving them is highly prized. If we know that a proposition is metaphysically possible or metaphysically necessary, then, it is held, we know something very important about it. Now the foregoing theory leaves it somewhat puzzling why this should be so. This is because the theory represents propositions of types (i)–(vii) as merely stating necessary conditions of existence in other possible worlds. They do not state sufficient conditions of trans-world identity. This comes to the fore when we consider consequences of the implicit definition like this one:

(c) If George VI was the father of Elizabeth II, then it is necessary that he was her father.

In combination with the fact that George was indeed the father of Elizabeth, this tells us that it is a necessary condition of Elizabeth's existing in any other world *W* that George be her father in *W*, but it does not imply that Elizabeth exists in *W*. Nor does it imply that Elizabeth exists in any other alternative to the actual world. Why would we want a metaphysical framework which tells us that actual entities can't exist in other worlds unless certain conditions are satisfied, but provides no basis for saying that actual entities *do* exist in other worlds? It seems that the necessary conditions would be of little value unless there were also sufficient conditions. If this is right, then the foregoing implicit definition of metaphysical necessity is incomplete. In addition to the clauses cited above, an adequate implicit definition should contain clauses that state sufficient conditions of trans-world identity. I will not attempt to formulate principles of this sort here, and in fact, as a review of the literature will attest, it has proved quite difficult to come up with principles that can withstand scrutiny. (See, e.g. Mackie 2006.) But it seems that if the metaphysical modalities really do have the value that is usually claimed for them, such principles must exist, and must be built into our modal concepts.

I find this line of thought persuasive, and will therefore presuppose in the sequel that true propositions of forms (i)–(vii) owe their necessity to their being consequences of principles that state necessary and sufficient conditions of trans-world identity.

5. Nomological Modalities

Roughly speaking, a proposition is nomologically necessary if either it is metaphysically necessary or it follows from the laws of nature. As with conceptual necessity and metaphysical necessity, we can suppose that this new modal notion can be captured by an implicit definition. In this case, the definition will imply that all of the following are necessary: the laws of logic, the laws of concepts, the constitutive mathematical laws, any additional laws of mathematics that are required for natural science, all true propositions of types (i)–(vii), and the laws of nature. It will also contain the laws of an appropriate modal logic. Assuming that it is possible to be epistemically justified in believing *any* propositions

concerning nomological necessity, we can be justified in believing the clauses of an implicit definition that has these consequences simply in virtue of having the concept of nomological necessity in our repertoire. As in other cases, it is possible to form justified beliefs about the corresponding form of possibility by constrained conceiving, though now the constraints must include some substantive portions of mathematics and the laws of nature.

In addition to the categorical nomological modalities, there are also relative nomological modalities that pose additional metaphysical and epistemological questions. In considering questions of categorical possibility, we simply ask whether propositions are compatible with the laws of nature; but in considering questions of relative possibility, we ask whether propositions are compatible with the laws of nature *together with* certain assumptions about the conditions that prevail in a given context. It is categorically possible for a human being to run a mile in under four minutes, but it isn't possible for *me* to do so, *given* certain particular facts about my body-type and level of training. Generally speaking, when we are considering opportunities and risks, we are focusing on the nomological possibilities that are presented by specific contexts. That is, we are focusing on questions of the form, "What are the nomological possibilities confronting agent *A*, given that *A* is in a situation of type *S*?" Being able to answer questions of this sort is crucial to human welfare. But how do we arrive at such answers? How do we gain knowledge of relative possibility?

Abstractly considered, the task of determining what is nomologically possible relative to an agent *A* and a situation *S* involves constructing a model that satisfies the laws of nature, a description of *A*, and a description of *S*, and then asking whether scenarios of interest are also satisfied by that model. It seems likely, however, that nature has endowed us with practical heuristics for answering specific categories of questions about relative possibility. Thus, for example, instead of constructing a model and showing that it satisfies certain relevant laws of nature, it seems that our job is generally the easier one of searching memory for information about what happened in situations similar to the present one that we have encountered in the past. Information about these past outcomes is what makes it possible for us to draw conclusions about the opportunities and risks that are presented by the current situation. At a more fundamental level, as J. J. Gibson pointed out, we probably have certain perceptual modules that are dedicated to recognizing such basic relative possibilities (or "affordances") as graspability, reachability, and climbability (Gibson 1979). We know that heuristics of both these types are reliable, for the plans and decisions that they support are often successful.

In addition to providing a metaphysical framework that makes it possible for questions about opportunities and risks to arise and be addressed, the nomological modalities also provide a framework for counterfactual reasoning, and for the enterprise of assessing counterfactual conditionals for truth or falsity. Thus, when we ask what would have happened if certain circumstances had been different than they actually were, we are in effect asking what would have happened in a situation as close as possible to the actual situation, but different from it in the given respects, had evolved in accordance with the actual laws of nature.

Questions of this type are important to us for a variety of reasons, including especially the fact that such questions enable us to figure out the contributions that various causal factors make to outcomes (by determining what an outcome would have been if a certain factor had not been present), and the fact that they enable us to assign responsibility to

agents for the consequences of their actions (by asking how things would have turned out if an agent had not performed an action, or had performed it in a different way). How do we answer questions concerning counterfactual processes and counterfactual courses of action? It seems that certain cognitive modules are dedicated to counterfactual reasoning, and that science is beginning to understand the ways in which such modules work. (See, e.g. Byrne 2007.) These modules must of course be included in any catalog of the sources of modal knowledge. (As we will see in the Appendix, it is sometimes maintained that the modules in question can be used to attain knowledge of the metaphysical modalities as well as knowledge of the nomological modalities. Here, however, I mean to be focusing on the simpler view that is concerned only with the latter sort of knowledge.)

It is sometimes maintained that the metaphysical modalities have little cognitive value beyond that which is possessed by the nomological modalities, and there are even philosophers who favor the more extreme view that the former coincide with the latter (Edgington 2004). It is understandable that people would be tempted by these ideas, since the practical importance of modality is largely due to its role in planning, decision-making, causal analysis, and assigning responsibility to agents, and as we have just seen, the modal requirements of these endeavors seem to be fully met by the nomological modalities. Thus, we again face the question of whether the metaphysical modalities are genuinely important, and also the question of why knowledge of them should be prized. Fortunately, reflection shows that these questions admit of reasonably simple answers: Even though the metaphysical modalitites have little practical value, they offer a number of fairly straightforward theoretical advantages. Thus, in the first place, we often find it theoretically interesting to consider what the world would be like if the laws of nature were different in one or more respects. This enables us to assess the degree to which the laws of nature are interdependent, and the degree to which some are more fundamental than others. We would of course lose these benefits if we focused narrowly on nomological possibility. The metaphysical modalities also provide a framework in which it is possible to frame definitions of important metaphysical concepts, such as supervenience, realization, and reduction. Among other things, these notions add considerably to our understanding of relations between actual objects, properties, and states of affairs. A third advantage is that the metaphysical modalities afford a basis for intuitively plausible explanations of key concepts from semantics, such as the notion of a truth condition.

6. Cartesian Modal Arguments

As is well known, Descartes thought it possible to establish that his mind and body were distinct by an argument based on a claim about the relationship between conceivability and possibility (Descartes 1984, p. 54). Abstracting from some details that need not concern us here, his argument comes to this:

> A. *First premise*: We are able to conceive of Descartes's mind existing without being accompanied by Descartes's body. (Justification: Descartes's evil genius scenario)
>
> *Second premise*: If we are able to conceive of Descartes's mind existing without being accompanied by Descartes's body, then it is objectively possible for this to occur.

Lemma: It is objectively possible for Descartes's mind to exist without being accompanied by Descartes's body.

Third premise: If it's possible for *x* to exist without being accompanied by *y*, then *x* is distinct from *y*.*Conclusion*: Descartes's mind is distinct from Descartes's body.

Because they tend to doubt the Cartesian doctrine that the mind is a substance, on a par with the body, contemporary philosophers tend to reject this argument; but they often accept, or at least show respect for, closely related arguments that are concerned with qualitative mental properties, like pain and the sensation of heat. Continuing to prescind from irrelevant details, we can represent the view of these philosophers by saying that they see merit in arguments like this one:

B. *First premise*: We are able to conceive of pains existing without being accompanied by any physical events.

Second premise: If we are able to conceive of pains existing without being accompanied by any physical events, then it is objectively possible for this to occur.

First Lemma: It is objectively possible for pains to exist without being accompanied by physical events.

Second Lemma: It is objectively possible for instances of the property *pain* to exist without being accompanied by instances of any physical property.

Third premise: If it is objectively possible for instances of a property *P* to exist without being accompanied by instances of a property *Q*, then *P* is not identical with *Q*.

Conclusion: The property *pain* is not identical with any physical property.

As the reader will have noticed, neither of these arguments makes a general claim about the relationship between conceivability and possibility. None of them asserts that conceivability is always or even usually a reliable source of modal knowledge. Instead they make comparatively modest claims concerning conceiving that are restricted to mental phenomena. Those who favor the arguments would maintain, I believe, that this restriction protects them from the strictures concerning conceivability that have been put forward in earlier sections. I will be concerned in the present section to evaluate this view.

I will focus exclusively on arguments like B. In discussing B., I will understand it to be concerned with metaphysical possibility. This is in keeping with the intentions of contemporary advocates of such arguments.

Now at first sight, B. appears to have a substantial flaw. As we saw in section 4, Kripke and Putnam put forward persuasive reasons for thinking that knowledge of the essential properties of empirical kinds is a posteriori. More specifically, they maintained that natural science provides our only reliable access to the essential properties of empirical kinds. Now B. is concerned with pain, and on the face of it, pain is an empirical kind. Our knowledge of pain comes from experience. But B. presupposes that it is possible to grasp the essential nature of pain simply by deploying the faculty of conception. Thus, according to the second premise of B., conception alone suffices to establish that pain can exist without being accompanied by physical phenomena, and this implies that conception alone can be used to show that physical phenomena are not implicated in any way in the essential nature of pain. Hence, the second premise seems to be flatly opposed to the lessons we learned from Kripke and Putnam. Either it is wrong or Kripke and Putnam were mistaken.

Although this objection was stated with considerable force by Kripke himself, he did not see it as fatal, and in fact he formulated an ingenious reply to it. The central thesis of his reply is that pain is a special case. Unlike other empirical kinds, qualitative empirical kinds like pain are known to us directly. We are immediately acquainted with them. As a result of this privileged access, we are intimately familiar with their essential natures. In the case of pain, for example, acquaintance provides us with an exhaustive knowledge of what it *is*. Further, in conceiving of pain, we rely on concepts that are based squarely on immediate acquaintance, and that therefore reflect all of the features of pain that acquaintance reveals. It follows that in conceiving of situations involving pain, our conceiving is constrained by the essential nature of pain. But this means that when we conceive of a situation involving pain, that situation really is an objective possibility. In other words, the second premise of B. is acceptable.

In formulating this line of thought, Kripke relied heavily on the notion of *picking pain out*. Thus, for example, he contrasted the way in which we pick heat out with the way in which we pick pain out. Heat, he maintained, is picked out as the external phenomenon that causes the sensation of heat. In other words, it is picked out as the external phenomenon that gives rise to a certain sort of appearance. But there is no appearance/reality distinction in the case of pain: "in the case of a mental phenomenon, there is no 'appearance' beyond the mental phenomenon itself" (Kripke 1980, p. 154). Hence, pain is picked out simply as an experience of a certain sort—specifically, as the experience of pain: "Pain, on the other hand, is not picked out by one of its accidental properties; rather it is picked out by the property of being pain itself, by its immediate phenomenological quality" (Kripke 1980, p. 152). This immediate phenomenological quality is the essence of pain: to grasp it is to grasp pain's essential nature, and to grasp that nature in its entirety.

What should we make of this set of views? It seems true enough that we are directly acquainted with pain, if this means only that awareness of pain is not mediated by awareness of something else. However, Kripke seems to think that if we are directly acquainted with pain, in the given sense, then we somehow have full access to the essential nature of pain. That is to say, he thinks that our epistemic relationship to pain must allow us full access to the modal facts about pain—and in particular, that it must allow us to answer such questions as whether it is possible for pain to exist without being accompanied by any physical kind. When we reflect, however, we see that the fact that a phenomenon is self-presenting does not imply that we have full access to its essential nature. To say that we are directly acquainted with a property *P* is simply to say that there is no other property *Q* such that we are aware of *P* in virtue of being aware of *Q*. But even if we are directly acquainted with a property in this sense, it can still be true that we do not have epistemic access to it that enables us to fully grasp its essential nature. For it can still be true that our awareness of *P* constitutively involves a representation of *P*, and it may be that this representation fails to fully reveal the essential properties of *P*.

In other words, Kripke appears to be presupposing the following principle:

> If we have *direct* epistemic access to pain, then when we imagine a possible situation that contains a pain, we automatically have *full* access to the essential nature of the pain we are imagining. Hence, since nothing physical presents itself to us when we are imagining the pain, we can be sure that we are not imagining anything physical in virtue of imagining the pain.

This principle claims that there is a close relation between *direct* access to the essential nature of a pain and *complete* access to its essential nature. It is far from obvious that this

claim is true. Indeed, it is closely related to a presupposition of Descartes's original argument that was challenged by Arnauld. (In his original version of the argument, Descartes took it as a premise that if we are able to conceive of *X* clearly and distinctly, then *X* is objectively possible. Arnauld objected to this principle on the grounds that our conception of *X* might fail to represent all of *X*'s essential properties. His example involved conceiving of a right triangle. We are able to conceive of a right triangle without conceiving of it as satisfying the Pythagorean condition that the square of the hypotenuse is identical to the sum of the squares of the sides, but it isn't objectively possible for a right triangle to exist without satisfying this condition. Arnauld 1984, pp. 139–42.). Descartes saw the force of this criticism and struggled to cope with it, but few readers have found his response satisfactory.)

To make this more concrete, suppose for a moment that pain can be identified with a type of bodily disturbance, and that awareness of pain involves a perceptual representation of a disturbance of the given type. On this view, it is possible to draw a Kantian appearance/reality distinction with respect to awareness of pain: there is pain as it is represented perceptually, and pain as it exists in itself. Moreover, as Kant would insist, it may well be the case that the appearance of pain fails to do justice to the reality. This is true even if the representation that registers pain represents it *directly*. Thus, suppose that pain is in fact represented directly by some representation *R*—suppose that there is no property *P* such that *R* represents pain by virtue of representing *P*. It could still be true that *R* fails to provide awareness of all of the essential properties of pain, because it could still be true that *R* lacks the right internal structure to articulate those properties. What is in fact a complicated physical structure could appear, from the perspective of someone whose awareness of pain is determined by *R*, as a simple, undifferentiated quality.

I am not claiming at the moment that pain *is* a bodily disturbance, or that our access to pain *is* perceptual, though I believe there is abundant evidence that these propositions are true (Hill 2009). I am just claiming that they are epistemic possibilities, and that they therefore must be ruled out before it can be maintained with confidence that our epistemic access to pain puts us in touch with the essence of pain. One cannot defend the latter view simply by pointing out that pain is self-presenting, or that we are directly aware of pain. But that is what Kripke tries to do.

It may be useful to state this point in a different way. For an a priori test to be reliable in the case of "Pain is not identical with any physical kind," it would have to be true that we have an a priori grasp of the *full* essential nature of pain. But there are credible stories about our epistemic access to pain which deny that this is so. In particular, a story which claims that awareness of pain consists in entertaining a perceptual representation of bodily damage may deny it. Before we can use conceivability to determine whether the given sentence is metaphysically possible, therefore, it would be necessary to rule out perceptualist accounts of awareness of pain on independent grounds. Since it is very unlikely that this can be done a priori, it is a mistake to think that it is possible to test the given sentence for metaphysical possibility in a purely a priori fashion. Unconstrained conceivability fails as a guide to objective possibility, even in the very special case involving qualitative mental states.

Cartesian modal arguments face additional difficulties. For example, they presuppose the erroneous thesis that it is possible for mere conception to afford us an exhaustive grasp

of the essential natures of physical phenomena. But I think enough has been said to show that in their present form, at least, such arguments are hopelessly simplistic.

7. Conclusion

I have claimed that various conceptions of necessity are presented to us by implicit definitions, and that these definitions constrain the use of conceivability in testing for the corresponding forms of possibility. There are alternative ways of approaching these topics that I have not considered here, but reflection shows, I believe, that the present approach provides the most unified and least problematic account of the various notions of necessity, and also the most plausible account of modal knowledge. I illustrate this claim in the following appendix.

APPENDIX

Can Necessity be Reduced to the Subjunctive Conditional?

In this Appendix I will say something about an alternative proposal concerning metaphysical necessity that has seemed right to me at various times in the past, and that appears to be enjoying a modest surge in popularity at present. According to the proposal I have in mind, metaphysical necessity can be defined in terms of the subjunctive conditional. There are three different forms that such a definition might take:

(a) It is metaphysically necessary that $P =_{df}$ if it were true that not-*P*, then it would also be true that *P*.
(b) It is metaphysically necessary that $P =_{df}$ it would be true that *P* no matter what else was true.
(c) It is metaphysically necessary that $P =_{df}$ if it were true that not-*P*, then a contradiction would also be true.

Lewis focused on (a) in *Counterfactuals* (Lewis 1973); Williamson and I proposed versions of (b) several decades later (Williamson 2005, Hill 2006); and in his recent work, Williamson has tended to emphasize (c) (Williamson 2007).

To appreciate the plausibility of these proposals, recall the standard view concerning the truth conditions of subjunctive conditionals:

> (*) A subjunctive conditional "If it were the case that *A*, then it would be the case that *C*" is true at a metaphysically possible world *W* just in case either (i) *A* is not true at any metaphysically possible world, or (ii) there is a metaphysically possible world at which *A* and *C* are true that is closer to *W* than any metaphysically possible world at which *A* and not-*C* are true.

On the assumption that (*) is at least roughly correct, it is possible to argue for version (a) of the theory as follows: "Suppose that *P* is metaphysically necessary. Then condition (i) of (*) is satisfied, because *not-P* is false at every metaphysically possible world. Given (*), it follows

that 'If it were true that not-*P*, then it would also be true that *P*' is true. Suppose now that this conditional is true. Given (*), it follows that either (i) *not-P* is not true at any metaphysically possible world, or (ii) there is a metaphysically possible world at which *not-P* and *P* are both true. Now if (i) obtains, *P* is metaphysically necessary. But (i) must obtain, since (ii) is absurd. Hence, *P* is necessary." It is possible to give closely related arguments for proposals (b) and (c).

The subjunctive theory of metaphysical necessity is appealing for several reasons, one being that it helps to explain the perceived *importance* of metaphysical necessity. Since the subjunctive conditional plays a role in a variety of important cognitive processes, notions that can be explained in terms of it have a reflected glow. Another advantage is that the theory holds out the promise of being able to explain knowledge of the metaphysical modalities in a uniquely appealing way. We know that there are powerful cognitive modules that process subjunctive conditionals, for it is a datum that we are capable of evaluating subjunctives, even if they are quite complicated, across a wide range of contexts. If the metaphysical modalities are reducible to subjunctive conditionals, we could enlist these modules in explaining how we acquire knowledge of the metaphysical modalities. Or so it can seem.

Unfortunately, however, these virtues of the subjunctive theory are balanced by difficulties. A particularly serious difficulty arises from clause (i) of (*). As is illustrated by the line of thought two paragraphs back, this clause will play a crucial role in any defense of the theory. Now in effect, (i) claims that all subjunctive conditionals with impossible antecedents are vacuously true. There are things that can be said in defense of this claim, but it is put into doubt by pairs of conditionals like the following, all of which have metaphysically impossible antecedents:

(1) If Obama had had different parents, he would have had different DNA.
(1') If Obama had had different parents, he would have been 6 inches tall.
(2) If a higher grade of cement had been used in the construction of Murphy Tunnel, the tunnel wouldn't have collapsed.
(2') If a higher grade of cement had been used in the construction of Murphy Tunnel, the tunnel would have been five miles long.
(3) If classical logic was false, the problem would lie with the law of the excluded middle.
(3') If classical logic was false, the problem would lie with modus ponens.

Propositions (1)–(3) strike us as true, or at least as reasonable things to say; but it would be wrong to say that they strike us as *vacuously* true, for if they did, (1')–(3') would also strike us as true, whereas they clearly strike us as false. Examples of this sort suggest that we should explore alternative ways of explaining subjunctives. More particularly, they suggest that we should do away with clause (i) of (*) and reformulate clause (ii) in such a way that it is based on a system of worlds that includes impossible worlds, including worlds in which Obama had different biological parents, worlds in which the Murphy Tunnel was constructed of different material at its origin, and worlds in which classical logic is false. But of course, if we base our account of subjunctive conditionals on a system that includes impossible worlds, we will not be able to explain metaphysical necessity in terms of subjunctives, for it will not be true that metaphysically necessary propositions are true in all of the relevant worlds. Where *A* is any proposition that counts intuitively as metaphysically necessary, there will be members of the relevant space of worlds in which *A* is false.

There is another way of dealing with the problem posed by pairs of subjunctives like the ones under consideration, but it is somewhat Byzantine compared to the solution that

invokes impossible worlds, for it commits us to a complex theory of the intentions of speakers who assert propositions like (1), (2), and (3). On this view, while (1) is vacuously true, speakers who assert it intend to communicate a quite different proposition, the proposition that Obama's biological parents are causally responsible for his DNA. Equally, speakers who assert (2) intend to communicate the belief that the inferior quality of the cement was causally responsible for the collapse of the tunnel, and speakers who assert (3) intend to communicate the belief that the law of the excluded middle is less obviously trustworthy than the other laws of classical logic. More generally, the thesis here is the Gricean idea that speakers recognize the triviality of subjunctive conditionals with impossible antecedents, and exploit that triviality in communicating more interesting propositions (Grice 1975). According to this idea, they assume that their conversational partners will also be aware of the triviality of the propositions, and that, being disinclined to believe that the speakers intend to make trivial claims, their partners will be moved by this perception to consider other propositions that the speakers may have intended to communicate, fixing finally on the ones that the speakers actually have in mind.

These claims have a certain plausibility, and at the end of the day, it may turn out that Gricean ideas provide an acceptable solution to the present problem. It should be noted, however, that the Gricean account of propositions like (1)–(3) is considerably more complex than the account that follows from the preceding theory, according to which the truth conditions of such propositions are best explained in terms of impossible worlds. An advocate of the preceding theory can say that a speaker will assert (1) just in case he or she believes that the truth condition of (1) is satisfied, but an advocate of a Gricean approach must say that asserting (1) is a complex affair—an affair that involves determining that the truth condition (1) is trivially satisfied, but that also involves fixing on another proposition that is true and that it would be useful to convey, and working out that the speaker's conversational partner will be able to recognize his or her intention to communicate that other proposition. This additional complexity appears to be a liability, for as far as I can see, there is neither introspective nor experimental evidence that we actually have the complex Gricean intentions that the second theory attributes to us. To be sure, if a speaker were to assert (1), she would no doubt expect her conversational partner to infer that Obama's parents are causally responsible for his DNA. But she would expect this because the proposition is a natural inference from (1) itself, not because she thinks that she can best get this proposition across by saying something too trivial to be taken at face value. In other cases, there is introspective evidence that Gricean intentions underlie assertions of propositions. A good illustration is Grice's example of the professor who writes a letter of recommendation saying only that a certain job candidate has excellent handwriting. We all know that if we were to write such a letter, we would be counting on the recipient to notice the triviality of our claim, and to infer from it that we do not have much confidence in the candidate. It would be our intention to damn the candidate with faint praise. But it is not at all obvious that we are guided by similar intentions in asserting propositions like (1)–(3).

A further point is that while the Gricean account presupposes that it is common knowledge that subjunctives with impossible antecedents are vacuously correct, this is actually quite far from the truth. If it was common knowledge that such subjunctives are vacuous, we would not be so perplexed by the problem that is posed by propositions like (1)–(3). Indeed, the problem of how best to interpret them would not arise. Even Lewis, who was on the whole a stalwart champion of (*), saw the doctrine of vacuous truth as a conjecture, acknowledging that the reasons in favor of it are "less than decisive" (Lewis 1973, p. 25).

I observe finally that the claim of the subjunctive theory to provide us with a uniquely appealing account of knowledge of metaphysical necessity is spurious. According to the theory, we come to know that a proposition *P* is metaphysically necessary by coming to know that a subjunctive conditional is true—say, to invoke definition (b), by coming to know the conditional "*P* would be true if any proposition were true." But how do we arrive at knowledge of the conditional? Presumably by showing that *P* can be obtained as the conclusion of a subjunctive argument no matter what proposition is taken as premise. But it would be impossible to show this unless there were a rule authorizing us to assert *P* at any stage in the construction of a subjunctive argument. That is, there would have to be a rule telling us that it is permissible to add *P* as a line in any piece of subjunctive reasoning. Suppose, for example, that *P* is the proposition that George VI is the father of Elizabeth II. How could it be true that this proposition is derivable from any premise we might adopt in the course of subjunctive reasoning? Reflection shows that this could be true only if there is a rule which says that if *x* is a biological parent of *y*, then it is permissible to assert a proposition to that effect at any point in any stretch of subjunctive reasoning. Now there is very little difference between a rule of this sort and a proposition of the following form:

($) If *x* is a biological parent of *y*, then for any proposition *P*, if it were the case that *P*, then it would be the case that *x* is a biological parent of *y*.

In adopting the rule, one is in effect adopting ($) as an axiom. So it is natural and appropriate to think of the subjunctive theory of necessity as committed to a set of propositions like ($). But this means that the subjunctive theory has essentially the same structure as the implicit definition theory that we considered in section 4, for the heart of that theory is a set propositions like (#):

(#) If *x* is a parent of *y*, then it is metaphysically necessary that *x* is a parent of *y*.

Thus, while the subjunctive theory takes a range of propositions like ($) to be constitutive of subjunctive reasoning, and therefore, to be constitutive of the subjunctive conditional itself, the implicit definition theory takes the corresponding range of propositions like (#) to be constitutive of metaphysical necessity. There is a deep structural parallel between the two theories. And by the same token, there is a deep parallel between the ways in which the two theories explain knowledge of metaphysical necessity. The subjunctive theory explains it in terms of our grasp of definition (b) (or one of the equivalent definitions) and our acceptance of a host of propositions like ($), and the implicit definition theory explains it in terms of our acceptance of the corresponding host of propositions like (#). If we evaluate the two theories simply on the basis of their explanation of modal knowledge, it seems that there is very little basis for giving preference to one over the other. It is wrong to suppose that the subjunctive theory has an explanatory edge.

In this most recent part of the discussion I have been assuming for the sake of argument that the subjunctive theory of necessity is viable. But of course, the arguments of earlier paragraphs call this assumption into question. One might have thought that there is at least reason to *hope* that the Gricean defense of the subjunctive theory can be made to work, because the subjunctive theory offers an explanation of modal knowledge that is uniquely privileged. The last paragraph is meant to show that this thought is misguided.

References

Arnauld, Antoine. (1984). *Fourth Set of Objections*, in John Cottingham, Robert Stoothoff, and Dugald Murdoch (eds.), *The Philosophical Writings of Descartes*, Vol. II. Cambridge: Cambridge University Press, pp. 138–53.

Boghossian, Paul. (1997). "Analyticity," in Bob Hale and Crispin Wright (eds.), *A Companion to the Philosophy of Language*. Oxford: Blackwell Publishers Ltd., pp. 331–68.

Byrne, Ruth M. J. (2007). *The Rational Imagination*. Cambridge, MA: MIT Press.

Descartes, Rene. (1984). *Meditations on First Philosophy*, in John Cottingham, Robert Stoothoff, and Dugald Murdoch (eds.), *The Philosophical Writings of Descartes*, Vol. II. Cambridge: Cambridge University Press, pp. 3–62.

Edgington, Dorothy. (2004). "Two Kinds of Possibility," *Proceedings of the Aristotelian Society Supplementary Volume,* vol. 78, pp. 1–22.

Gelman, Susan. (2003). *The Essential Child: Origins of Essentialism in Everyday Thought*. Oxford: Oxford University Press.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

Grice, H. P. (1975). "Logic and Conversation," in Donald Davidson and Gilbert Harman (eds.), *The Logic of Grammar*. Encino, CA: Dickenson, pp. 64–75.

Hill, Christopher S. (2006). "Modality, Modal Epistemology, and the Metaphysics of Consciousness," in Shaun Nichols (ed.), *The Architecture of the Imagination*. Oxford: Oxford University Press, pp. 205–35.

Hill, Christopher S. (2009). *Consciousness*. Cambridge: Cambridge University Press.

Keil, Frank. (1989). *Concepts, Kinds, and Cognitive Development*. Cambridge, MA: MIT Press.

Kripke, Saul. (1980). *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Kripke, Saul. (2011). "Identity and Necessity," in Kripke (ed.), *Philosophical Troubles*. Oxford: Oxford University Press, pp. 1–26.

Lewis, David. (1973). *Counterfactuals*. Oxford: Blackwell Publishing.

Mackie, Penelope. (2006). "Transworld Identity," in *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/identity-transworld/. Accessed September 20, 2015.

Putnam, Hilary. (1975). "The Meaning of 'Meaning'," in *Putnam, Mind, Language and Reality* (ed.), Cambridge: Cambridge University Press, pp. 215–71.

Sidelle, Alan. (1989). *Necessity,Essence, and Individuation*. Ithaca, NY: Cornell University Press.

Williamson, Timothy. (2005). "Armchair Metaphysics, Metaphysical Modality and Counterfactual Thinking," *Proceedings of the Aristotelian Society* 105, pp. 213–35.

Williamson, Timothy. (2007). *The Philosophy of Philosophy*. Oxford: Blackwell Publishing.

CHAPTER 19

PHILOSOPHICAL HEURISTICS AND PHILOSOPHICAL METHODOLOGY

ALAN HÁJEK

1. Introduction

Metaphilosophy is all the rage nowadays.[1] Philosophers are becoming increasingly self-conscious about their methodology, as this volume showcases. And rightly so—it is part of our job description to put thinking under the microscope, and that obviously should include our *philosophical* thinking. Much as we want philosophers of physics, of biology, of mathematics, and so on to scrutinize those respective disciplines, so we want philosophers of philosophy to scrutinize *ours*.

When I look at the ways that much philosophy gets done, I see certain recurring patterns of thought. The best philosophers repeatedly deploy a wealth of philosophical techniques. I see them used in papers, books, talks, Q & A sessions, and philosophical discussion more broadly. Since I was a graduate student I have observed them, and when they have seemed especially fecund to me, I have recorded them on an ongoing list and I have tried to internalize them. After all, they are part of the philosopher's intellectual armoury, much as logic is. We may call them *philosophical heuristics*.

But unlike logic, they are not studied or taught. We just *use* them, but we typically do not do so self-consciously; indeed, we may often use them unconsciously, unaware that we are even doing so. Despite the importance of such heuristics to philosophical methodology,

[1] Thanks especially to John Barker, Paul Bartha, Bert Baumgaertner, Sharon Berry, Rachael Briggs, Mark Colyvan, Tamar Szabó Gendler, Patrick Greenough, Robbie Hirsch, Yoaav Isaacs, Ben Levinstein, Hanti Lin, Peter Menzies, Bernard Molyneux, Daniel Nolan, Wolfgang Schwarz, Michael Simpson, Nick Smith, and to audiences at the 2012 Australian Association of Philosophy conference and at Macquarie University for helpful discussion and/or comments on earlier drafts. Thanks also to Ralph Miles for editorial assistance.

they have been surprisingly neglected by philosophers. (Nozick 1993 is a notable exception, but even his discussion of some of them is confined to just over eight pages.)

Practitioners of various other disciplines have not been so remiss regarding their own heuristics. Mathematicians, for example, often teach their students heuristics for approaching problems, and there are numerous books (e.g. Pólya's 1957 classic *How to Solve It)* detailing them. Here's one that I learned from a mathematics professor: if you are struggling to prove something because it seems obvious, try reductio ad absurdum. Or again, music professors teach their students various 'rules' of harmony and counterpoint, which really are rules of thumb: avoid consecutive fifths and octaves, and so on. No mathematician or musician pretends that once you have mastered such heuristics, you will be the next Gauss or Beethoven, but it would be absurd to question their utility as strategies for guiding one's thinking or creativity.

Almost any complex activity has its heuristics—cooking, rock-climbing, photography, gymnastics, playing the xylophone, salsa dancing, taxi driving, taxidermy (I assume) … It would be odd if philosophy, one of the most complex activities of which we are capable, *didn't* have heuristics. And it obviously does—hundreds that I've identified so far. Some are *exclusively* philosophical heuristics, while others are more general-purpose heuristics for conducting one's thoughts that one finds employed elsewhere. And still others are not so much 'heuristics' in a narrow sense, but fruitful patterns of thinking. Despite this chapter's title, I won't put much weight on either the word 'philosophical' or 'heuristics'. I am happy to include philosophically fertile strategies, however we might best label them.[2]

It would be strange counsel to remain silent about them: [whispered] 'Yes, they're out there (unfortunately), but don't tell anyone else about them!' And yet I have encountered some critics who view my project with the sort of suspicion visited upon a magician who breaks ranks with his fellow magicians by revealing some of the secrets of their trade. Well, I'll let you in on a few of the secrets of *our* trade. But then, I bet that you tacitly know some or all of them anyway; I'm just making them explicit.

I have already presented a first batch of heuristics in my (2014). There I discussed such techniques as:

- *Check extreme cases, and near-extreme cases.* These are especially promising places to look first for counterexamples, and doing so reduces your search space.
- *Death by diagonalization: reflexivity/self-reference.* Plug into a function itself as its own argument, and more generally, appeal to self-referential cases. It worked for Cantor, Russell, and Gödel; it may work for you.
- *Self-undermining views.* Relatedly, check whether a philosophical position or argument that falls in the very domain that it purports to cover treats itself consistently.
- *Begetting new arguments out of old.* Good arguments are often easily transformed from one domain to another. Exploit analogies between, e.g. space, time, and modality; and between rationality and morality.
- *Trial and error.* When ingenuity fails you, sometimes you can just run systematically through the relevant cases until you find one that meets your needs.

[2] For example, Kahneman and Tversky (1982) use the word 'heuristic' in a somewhat different sense from mine.

I stressed that they are not sure-fire recipes to philosophizing—there are none—but rather defeasible guides that tend to work.

Here comes a new batch. I will focus on a cluster of heuristics that have some interesting family resemblances. Indeed, *resemblance* is a theme that unites several of them. For instance, there may be multiple candidates for some role (a concept, a choice, a solution to a problem, or what have you), which are very *similar* in relevant respects. As a result, they appear to play the role equally well. How should we respond to that? I canvas various heuristics that are motivated by or that exploit this problem, which will then fan out to other related heuristics.

Away we go.

1.1 See Definite Descriptions in Neon Lights

A philosophical thesis or an analysis that involves a definite description '… *the F* … ' typically presupposes that there is *exactly one F*. Ask:

i) Are there, or could there be, *multiple F*'s?
ii) Are there, or could there be, *no F*'s?

In short, check to see if there are *any* or *many F*'s. The hard task of looking for problem cases has been broken down into two easier sub-tasks. i) and ii) correspond to tests for *uniqueness* and *existence*, respectively, familiar in mathematics.

Examples:

1.1.1 *Counterfactuals*

Stalnaker (1968) analyses counterfactuals[3] along the following lines:

> if it were the case that *X*, it would be the case that *Y*' is true
>
> iff
>
> *Y* is true at THE nearest *X*-world.

(Since I do not have a neon-lights font at my disposal, the capital letters will have to do.) The 'THE' indicates the assumption that for any *X*, there is a *unique* nearest *X*-world. Lewis (1973a) has two main objections:

i) There may be *multiple* nearest *X*-worlds. 'If Bizet and Verdi were compatriots, then . . .' Would they both be Italian, or both be French? It seems that nothing favours one nationality over the other. So there are at least *two* nearest worlds in which they are compatriots. More generally, Stalnaker's analysis appears to founder on cases in which there are ties for the nearest *X*-world***s***.
ii) There may be *no* nearest *X*-worlds. 'If I were taller than 7 feet, then …' How tall would I be? 7 feet 1 inch? 7 feet ½ inch? 7 feet ¼ inch? . . .' We apparently have an infinite

[3] In fact, his analysis is of conditionals in general, but Lewis's reply, to which I turn, concerns counterfactuals more specifically.

sequence of ever-closer worlds in which I am taller than 7 feet, with none closest. More generally, Stalnaker's analysis appears to founder on cases in which there are *no* nearest *X*-worlds.

1.1.2 *The Problem of Evil*

A well-known argument against various forms of theism goes like this:

> If an omnipotent, omniscient, omnibenevolent God existed, then such a God would have created THE best of all possible worlds. But the actual world is not the best of all possible worlds; therefore such a God does not exist.

i) Perhaps there are many worlds tied for first place in the relevant 'goodness' ordering. In that case the argument needs to be recast: God would have created (at least?) one of them, and the actual world is not one of them.
ii) Perhaps worlds get better and better without end—none best. Imagine an infinite sequence of worlds in which there are successively more happy people, or more happy rabbits … , each world better than its predecessor.

1.1.3 *Functionalism*

Functionalist accounts of theoretical and folk-theoretical terms often invoke locutions such as 'the *X* role' (neon lights), or better still for my purposes, 'the occupant of the *X* role' (double neon lights). This presupposes that there is exactly one such role, and exactly one such occupant. Functionalists speak, for instance, of 'the occupant of the pain role' (which is ritually taken to be the firing of c-fibres). Lycan (2009) argues that the presupposition of a single occupant begs the question against the dualist view according to which pain behaviour is causally overdetermined, being caused both by physical neural events and non-physical pain events. One might also question whether there is an occupant of the role at all; and whether there is a single pain role—for example, perhaps different folk theories assign it different roles.

So far, the examples have been familiar. Now, let me use the heuristic to make a less familiar point—to do some new philosophical work.

1.1.4 *Velocity as the Time-Derivative of Position*

Nearly every physics textbook analyses velocity as the time derivative of position:

$$v = dx / dt.$$

THE time derivative. Let's subject this analysis to our heuristic:

i) I see no problem with there being multiple time-derivatives of position. To be sure, there may be many time variables, corresponding to many frames of reference; but

in that case there are correspondingly many velocities, one for each time variable. The heuristic came up empty-handed here.

ii) I *do* see a problem with the time-derivative not existing. There are functions that are not differentiable everywhere; in fact, there are functions that are differentiable *nowhere*; in fact, there are functions that are *continuous everywhere and differentiable nowhere*. Weierstrass discovered such pathological 'saw-tooth' functions. Now think of a particle whose position function of time is such a function. This particle moves highly erratically, darting this way, then that, in a wild manner. Yet it traces a continuous path, so it is implausible that it goes in and out of existence. In short, we apparently have an example of *motion without velocity*. This may not be an outright counterexample to the physicists' analysis of velocity, but it is surely odd. The heuristic helps to lead us to a puzzling case.

In response to the problem of multiple candidates for *F* in the definite description 'the *F*', one may settle instead for the *indefinite* description, '*a(n) F*'. Then any of the candidates should do: all that matters is their existence, not their uniqueness. But we may still be able to mount our assault from the other side: even *indefinite* descriptions are in trouble when their existence presuppositions fail. This brings me to a closely related heuristic.

1.2 See Indefinite Descriptions in Neon Lights

A philosophical thesis, or an analysis, that involves an indefinite description '… *a(n) F* …' presupposes that there is *at least one F*. Ask whether there are, or could be, *no F*'s. All of the cases we have seen faced a challenge of this kind, so even replacing their definite descriptions with indefinite descriptions would not get them out of the woods.

1.3 Exposing a Definite or Indefinite Description by Paraphrase

Sometimes a problematic definite or indefinite description can come in disguise, and when it does, it is harder to see any neon lights, because one does not actually see the word 'the' or 'a(n)'. Still, it may be lurking in the background, to be revealed in its neon glory by paraphrase.

Example. To *maximize* a quantity means to achieve THE greatest amount of that quantity. The paraphrase makes the definite description explicit. But we can also speak of a function that has multiple maxima, many points at which it achieves its maximum value. Think of $y = \cos x$—while in a sense it has one maximum (namely, 1), in another sense it has many maxima (at $x = \pm 2n\pi$, $n = 0, 1, 2, \ldots$). In the latter sense, each of these points is A maximum, the indefinite description made explicit. The same can be said for *minimizing* a quantity, and *minima*, mutatis mutandis. But whether we speak of the single maximal/minimal value attained by a quantity, or the many ways in which it may be attained, one cannot maximize/minimize a quantity that has *no* maximum/minimum.

For instance, according to decision theory, rationality requires you to *maximize* expected utility. It does not seem problematic for decision theory if there are multiple ways to do so in a given situation. Consider Buridan's Ass, who can maximize expected utility by eating either of two equidistant hay bales; either way, it will achieve THE maximum amount of expected utility. On the other hand, cases in which there is an infinite sequence of actions of ever-greater expected utility are problematic. Pollock's (1983) 'Ever-better wine' provides an example: the longer you wait to open the bottle of wine, the better it gets. When should you open it? We can specify the case so that any time is too soon; yet never opening it is the worst option of all. There is no sense in which you can maximize expected utility here.

Similarly, a naïve version of consequentialism says that one is morally required to perform an action that has maximally good consequences (in some sense or other); but one cannot if there are actions that yield better and better consequences without end. Moving to a higher level of abstraction, some functions have no maximum at all—those that increase towards an asymptote without ever reaching it, or that have no bound, or whose range is an open set.

1.4 Indeterminacy

So far, we have looked at cases where 'the *F*' does not have a unique referent, because there is determinately more than one *F*, or determinately none; and cases where 'a(n) *F*' does not have a referent, because there are determinately none. A different kind of problem arises when it is *indeterminate* how many *F*'s there are, and in particular, it is indeterminate whether there is *one*, or indeterminate whether there is *at least one*. Similarly, the word 'the' suggests that there is a determinate answer to a relevant question, or a fact of the matter of the bearer of some relevant property. But there might be indeterminacy regarding these things.

Example. Philosophers of biology ask questions such as 'What is *the* function of the frog's eye?' And various answers seem reasonable: 'to detect flies', 'to detect dark spots', 'to detect food', and so on. But perhaps there is no determinate answer to the question—no determinate function of the frog's eye.

Example. What is *the* right thing to do when one faces a moral dilemma? Parfit (1984) argues that often it is indeterminate. It may even be indeterminate whether there is *a* right thing to do.

We will encounter indeterminacy again in our next heuristic, which has a number of points of contact with the previous ones.

2. Arbitrariness, and How to Respond When Faced with It

When a philosophical position has the form '… the *F* … ', or … a(n) *F* …' and there are multiple *F*'s, it may be permissible just to pick one of them. For example, if there are many best

of all possible worlds, maybe God can create any one of them. But perhaps choosing one *F* out of the many candidates will be *arbitrary* in an unacceptable way. God's creating one out of multiple worlds tied for first place in the goodness ordering would violate the principle of sufficient reason. For that reason, when Leibniz thought that God created *the* best of all possible worlds, he really meant *the*, not *a*!

Onwards to the next heuristic, or better, set of heuristics. They begin with problems arising out of arbitrariness, and then offer a number of ways of responding to these problems.

The sorites paradox furnishes a classic example:

> A grain of sand is not a heap; two grains of sand is not a heap; for all *n*, if *n* grains of sand is not a heap, then neither is *n*+1 grains; therefore there are no heaps.

Sure, you can arbitrarily stipulate that there is a particular grain of sand at which a heap suddenly comes into being—say, the seventeenth. But that's exactly the problem: it's arbitrary. Why that choice, rather than the sixteenth, or eighteenth, or other nearby choices?

Arbitrary choices are familiar from daily life. On what side of the road should we drive? This seems to be entirely arbitrary. Some other choices are not entirely arbitrary, but partly so. Where should we set speed limits, or the voting age, or the drinking age? Some settings would be far too high or far too low, but still there is a range of reasonable settings, and choosing one from the range is arbitrary. A sign of the arbitrariness is that different societies choose differently, and sometimes a given society will revise its settings over time. But we are typically not troubled by their arbitrariness, for it is *not* arbitrary that *some* setting be made, as opposed to none.

On the other hand, arbitrary choices are often regarded as fatal to philosophical positions that have to make them. The classical interpretation of probability, with its notorious principle of indifference, and Carnap's (1952) logical interpretation of probability, have been widely thought to be killed, or at least seriously wounded, by the apparently arbitrary choices that they force upon us—the suitable partition of 'equipossible' events in the former case, or the setting of an index of inductive caution, λ, in the latter.

More broadly, arbitrariness can be a sign of a flaw in a philosophical position. It is forced to make a choice; but why that choice, when it apparently could just as easily have made another one? The choice would seem to have no force—normative, or semantic, or otherwise. How could it be binding—rationally, or morally, or semantically, or what have you? And it would seem not to line up with a joint of nature, or of metaphysics, or of semantics, or what have you.

A good case study, which will showcase several techniques, involves Lewis's (1973a, 1994) analysis of *laws of nature*. Start with all of the true theories of the world. Some are very simple, but not informative—for example, the theory whose sole axiom is that everything is self-identical. Some are very informative, but not simple—for example, the collective (true) contents of *Wikipedia*. Some achieve a better balance of simplicity and informativeness than others. According to Lewis, the laws of nature are the theorems of the true theory of the world that best balances simplicity and informativeness—for short, *the* best system. But he acknowledges the possibility that there may be more than one reasonable way to trade off simplicity against informativeness. Different standards for balancing simplicity and informativeness may yield different theories as the winner of the Lewisian competition. What, then, are the laws? We could simply choose one set of standards, and insist that

it dictates what the laws are. But why that set, rather than another set? This choice threatens to be arbitrary.

More generally, a problem arises for a philosophical position when there are multiple candidates for some job description appealed to by that position, all apparently equally good, and choosing any one of them over the others seems arbitrary.

But there are various possible responses. I will classify them into three kinds:

1) Symmetry-breaking responses (playing favourites): all candidates are equal, but some are more equal than others.
2) Symmetry-preserving responses (even-handed): all candidates really are equal, but we can deal with that.
3) Hybrid responses: first some symmetry-breaking; then, symmetry-preserving among the candidates that remain.

2.1 Symmetry-Breaking Responses

2.1.1 *One of the Candidates is Salient, or Privileged*

The first response is to insist that one of the candidates stands out after all, so choosing it over the others is not arbitrary after all. I do not have an account of what makes a candidate salient or privileged, but I do have some rules of thumb. In daily life, when numbers are involved, natural numbers are usually salient. For example, no society imposes a drinking age of 18½ years old. Some natural numbers are especially salient—for example, 1 and multiples of powers of 10 often stand out from their neighbours. I bet you have never seen a speed limit of 103 km/h! In daily life and in philosophy, extremal cases are usually salient. So too are points of symmetry, and they can provide symmetry-breakers! In doing so, they can answer charges of arbitrariness. One candidate stands out: the symmetrically placed one. Consider our dividing a pie that you and I both want—if we don't agree on the 50/50 split, what could we agree on?

And in philosophy more specifically, one candidate for a job description may be privileged in virtue of being more fundamental than its rivals—for examples, its rivals are reducible to that candidate, or supervene on it, but not vice versa. Then it may be appropriate to thump the table in favour of that candidate. Or a candidate may be privileged, but not for any deep metaphysical reason. It may be so merely in virtue of contingent facts about our linguistic practices, for instance.

I take Lewis to have been sympathetic to a 'privileged candidate' response for his account of lawhood: that there are privileged standards for balancing simplicity and informativeness. We could begin to flesh out this idea—for example, by considering the lengths of descriptions of the world given by various competing axiomatized theories, couched in some canonical language in which all predicates correspond to natural properties. Some authors have similarly attempted to rehabilitate the principle of indifference by regarding certain partitions to be privileged—for instance, those that are maximally fine grained (Elga's 2004 'predicaments'), or that are invariant under certain fundamental transformations (Jaynes 2003).

While Lewis does not say in his writings what would *make* a balancing-standard privileged, still less what the privileged standard *is*, Elga and Jaynes offer more details about how privileged partitions are generated. One might offer the privileged-candidate response as a mere existential commitment—that there *is* a privileged candidate, and to leave it at that—but one might also do more to identify the candidate. The former approach might work especially well in tandem with *epistemicism* about the candidate: for some reason its identity is obscure to us or unknowable by us. (One hopes at least for more details about what *this reason* is, lest we be saddled with epistemicism about it too!) The poster child here is Williamson's (1994) approach to vagueness. According to it, there is a fact of the matter of which additional grain of sand turns a non-heap into a heap (thanks to our linguistic practices), but we can never know what grain it is.

2.1.2 *Go Subjectivist/Pragmatist*

The next response runs: '*You* get to choose the candidate that you want—it's *your* interests that you want to serve!' This response may be plausible when the relevant job description involves *subjects*, for example rational agents. Subjective Bayesians about probability, for example, are often untroubled by the seeming arbitrariness in the choice of priors. The response might also allow or explain no-fault disagreements between different subjects. But it is surely implausible for the laws of nature.

2.1.3 *Go Conventionalist*

'*We* get to stipulate the winning candidate; having done so, we agree on it thereafter.' (Note that arbitrariness of a choice was a key part of Lewis's (1969) definition of that choice being a matter of *convention*.) This response works for some of the laws of society, up to a point—for example, the side of the road on which one drives, the setting of speed limits, the voting age, and drinking age. Again, this response is not so promising for the laws of nature—we don't get to decide what they are. Some philosophical problems should not be farmed out to sociologists.

2.1.4 *'Nature is Kind'*

'The multiplicity problem is not really a problem, because however we reasonably make the arbitrary choice, there will be the same clear winner. While there might in principle be disagreement among the multiple best candidates for some job description, *in fact* such disagreement will not arise. Think of this as an empirical bet that nature is kind to us, and will see to it that all these candidates agree.' Lewis favoured this response regarding the laws of nature. He hoped that one theory would win the Lewisian competition so decisively that it would not matter exactly how simplicity and strength were traded off, within reasonable limits. This response is perhaps not so effective for an *analysis* of lawhood that is intended to hold in other possible worlds, including those in which nature is unkind; but perhaps our concept of lawhood would not apply in such a world.

2.2 Symmetry-Preserving Responses

2.2.1 *Pluralism*

'All of the candidates are right. Each provides a legitimate meaning, a reasonable explication of an inexact concept or a reasonable precisification of a vague concept.' This may be a good response when the arbitrariness at issue is purely semantic.

At the other end of the spectrum, we have eliminativism.

2.2.2 *Eliminativism*

'None of the candidates are right. The multiplicity of candidates serves to show that the original concept is incoherent, and should be eliminated.' This may be a good way to bring out problems with a concept that was already in dispute, as 'law of nature' is to some extent. But there is a danger of this response 'proving too much'. Most of our concepts are vague, and susceptible to sorites reasoning. We may be left with very few of them if we wield this response too enthusiastically. To be sure, however, even ordinary objects are under threat from such reasoning by eliminativists such as Unger (1979) and van Inwagen (1990).

Somewhere between the extremes of this response and the previous one, we have the following reponse:

2.2.3 *It's a Terminological Matter*

'It's not that any given candidate is right or wrong. Proponents of different candidates are merely taking different stands on a terminological issue.' Again, this may be a good response when the arbitrariness at issue is purely semantic. But of course the setting of the side of the road one drives on, speed limits, the drinking age, or voting age is not a terminological matter.

2.2.4 *It's Indeterminate*

'There is no fact of the matter of what the right candidate is. Rather, it is indeterminate.' This could be a good explanation of why the leading candidates are tied; or their being tied could explain why the matter is indeterminate.

2.2.5 *Supervaluate*

'What's true on *all* ways of making the arbitrary choice is determinately true. What's false on *all* ways is determinately false. Everything else is indeterminate.' According to Lewis (1973a), the laws are those theorems *common to all* of the candidates for the best theories. On this approach, as he later (1994) concedes, there is the threat that there will be little or no overlap between the best theories, in which case there will be few laws or none. We might hope that nature is at least a bit kind, guaranteeing a decent amount of overlap.

Recall the problem for Stalnaker's account of counterfactuals: that there may be ties for which antecedent-world is nearest. He later (1981) responds by supervaluating over the candidates.

2.2.6 *Subvaluate*

'What's true/false on *all* ways of making the arbitrary choice is determinately true/false. Everything else is true *and* false.' On this approach, there is the threat that there will be too *many* laws. And some of them may be highly disjunctive, piecing together regularities favoured by one best system with those favoured by another, and yielding something that is not simple by any standard.

Here, and in following pluralism by eliminativism, we encounter another mini-heuristic: *do the opposite*. Take some approach to a problem, and follow the opposite approach, dual approach, or complementary approach. (Suitably cautioned by my first heuristic, this assumes there is exactly one such approach!) For example, replace universal quantifiers by existential quantifiers, necessities by possibilities, conjunctions by disjunctions, or replace parameter values by values that are in some sense complementary. In this case, subvaluationism can be regarded as the opposite approach to supervaluationism.

And even here we face some arbitrariness, now at the meta-level. How do we justify supervaluating and subvaluating *over* other ways of meta-valuating, quantifying over the ways of choosing candidates? To be sure, supervaluating and subvaluating are *salient*, in virtue of being extreme, the two endpoints of a spectrum of ways of valuating: what's true on at least one way of making the choice; what's true on at least two ways; what's true on at least three ways; … what's true on all ways of making the choice. But salient too is 'majority-rules-valuating', which appeals to a point of symmetry, the midpoint: what's true on *more than half* of the ways of making the arbitrary choice is true; what's false on *more than half* of the ways is false.[4] (What's true on exactly half the ways we might treat as indeterminate, or we might treat as true and false.) Moreover, how are we to choose *between* supervaluating and subvaluating, when they are equally salient? *This* choice threatens to be arbitrary. A big debate begins here, turning on considerations that militate in favour of one approach or the other, and this is not the place—or even *a* place—to enter it.

2.3 Hybrid Responses

These responses combine some symmetry-breaking response (to cull some of the candidates while leaving others live) with some symmetry-preserving response (to treat those that remain even-handedly). One might insist, for instance, that *some* (but not all) of the candidates are privileged, then supervaluate over those that remain. Or one might consider all of the candidates that are subjectively chosen by some agent or other, and subvaluate

[4] If there are only finitely many candidates, it's clear what 'more than half' of them means—we just count them all and divide by 2! But if there are infinitely many candidates, it's less clear. We will need some sort of *measure* over the candidates. But which? Arbitrariness—dare I say it—looms!

over those. And so on. This yields many hybrid responses, mixing and matching techniques from the previous two categories.

I have presented many ways of responding to arbitrariness. Which should be used in a given case? Is that *arbitrary*?! How should we *respond*?! Different ways are appropriate for different cases, and I don't have a heuristic for deciding *that* (yet). I suggest that where possible, one should look to the symmetry-breaking strategies first, and only when they fail to leave just one candidate standing, look to the symmetry-preserving strategies over those that remain. But this still leaves open many possible responses, which will need to be handled on a case-by-case basis. Fear not—I have no aspirations to turn students into philosophical automata. There will always be an important place for good philosophical judgement.

So much for arbitrariness. It is continuous with the next set of closely related heuristics, which I will put under one big heading.

3. CONTINUITY

3.1 Continuity Reasoning

Let's revisit the sorites paradox from a different perspective. You can draw a line in the sand, so to speak, and claim that there is a particular grain at which a heap suddenly comes into being—say, the seventeenth. But it seems absurd that such a small change in one respect, the difference between 16 and 17 grains, should result in such a large difference in another respect, the non-existence or existence of a heap. Note that the problem here is not that of arbitrariness, justifying why the line should be drawn there rather than elsewhere. Rather, the problem is that *there shouldn't be a line at all*.

This is an example of *continuity reasoning*. Roughly, the pattern is that one variable is a function of another, and small changes in the former should lead to small changes in the latter. Such reasoning is often part of common sense. We are surprised, for instance, when we are told that we share 98.4 per cent of our DNA with chimpanzees. How can such a small change in genotype lead to such a large change in phenotype? More generally, discontinuities may induce some of the discomfort that arbitrariness may cause us. Why should cases that are similar in some relevant respect give rise to such dissimilarity in another respect? And where there is a discontinuous 'jump' in some function of interest or importance to us, both the placement and the size of the jump may seem disconcertingly arbitrary.

It should be stressed how much continuity reasoning underwrites inductive inference. We don't just think that the unobserved resembles the observed (in suitable respects): we also think that the *nearby* unobserved *closely* resembles the observed, and typically the more nearby, the closer the resemblance (other things being equal). The world mostly does not deliver abrupt changes; properties tend to change gradually over space and time. Consider how painting restoration, or computer programs for reducing noise on photographs, operate on this assumption: where there is information on only particular parts of a picture, the default assumption is that nearby parts will be the same, or similar. Induction would be stymied if things systematically underwent sudden jolts. And when it is, that just adds to our discomfort!

Continuity reasoning is also common and fertile in philosophy. Sider (2002) poses the problem that certain conceptions of Hell are incompatible with a traditional doctrine about God. According to these conceptions, after we die, we either go to Heaven or Hell; some of us go to Heaven and some go to Hell; Heaven is much better than Hell; and God decides who goes where. The concern is that any plausible criterion for his decision will admit of borderline cases. As a result, there will be some people who just make the cut and go to Heaven, and other very similar people who just miss out. Now Sider appeals to a continuity premise: 'the proportionality of justice prohibits very unequal treatment of persons who are very similar in relevant respects' (59). He argues that one cannot square such treatment with God being just.

One may argue in a similar way for vegetarianism being morally required. Clearly we should not be *omnivores*—for example, cannibalism is surely morally prohibited, and for that matter so is eating chimpanzees, most of us would agree.[5] If we are meat-eaters at all, we must have some criterion for deciding which animals it is permissible to eat and which not—as it might be, intelligence or sentience. Again this criterion will admit of borderline cases. Some animals will just make the cut (as it were) for being off limits for eating, and others will just miss out. And now comes the continuity premise: animals that are so similar to each other in relevant respects should not be given such disparate treatment. Conclusion: all animals are off limits.

We could argue along similar lines against the death penalty. Whatever criterion we use for drawing a line on the basis of severity of crime, beyond which perpetrators are executed, their punishment would be strikingly different from that of similar offenders just short of the line. Indeed, there is a whole class of 'slippery slope' arguments that have been deployed against euthanasia, abortion, gun control, and so on, which in each case appeal to a continuity intuition that similar cases should be treated similarly.

Of course, there are ways of fighting back. For starters, continuity reasoning can be run in both directions. We can equally 'show' that any collection of grains of sand, however small, is a heap by starting with a paradigm case of a heap, removing grains of sand one by one, and insisting that no single removal could turn a heap into a non-heap. The very same continuity premise that underwrites an argument for vegetarianism could just as well underwrite an argument for omnivorism. And eventually some crackpot is bound to appeal to continuity reasoning to argue for the far more widespread use of the death penalty, perhaps even supplanting fines for driving on the wrong side of the road. Beware of the *slippery slope* that the unbridled use of continuity reasoning could send us down! We can drive continuity reasoning in two directions—as we might say, left-to-right (→) *and* right-to-left (←). There is again the threat of arbitrariness: why should one direction be favoured over the other? Yet if we drive the reasoning in *both* directions, we will arrive at a contradiction (→←).

Moreover, continuity-based arguments must appeal to some sort of *metric*, at least loosely specified—some measure of 'distance' according to which we can judge roughly how close entities or cases are to each other. (It need not be numerical, but it must be more than merely a comparative ordering.) Disputes might arise over the choice of metric, and favouring one metric over another might appear to be arbitrary.

[5] I bracket extreme cases in which one's very survival is at stake, which may not be so clear.

And whatever the metric, sometimes discontinuities are acceptable, and even required. My favourite function is the *Alan Hájek* function. It is the function from people to {0, 1} that is my characteristic function: it assigns 1 to me, and 0 to everyone else. Offhand, this function should be discontinuous, whatever metric we impose to capture how similar people are to one another. It doesn't matter how close you are to me according to such a metric: you still get a 0, and only I get a 1!

3.2 Drawing Inspiration from the Mathematics of Continuous Functions

Let's look to mathematics for a better understanding of continuity, through its treatment of continuous functions. The informal definition of such a function is that small changes in its input value result in correspondingly small changes in its output value. More formally, the function *f(x)* is *continuous at c* if

$$\lim_{x \to c} f(x) = f(c).$$

This presupposes an underlying metric, as becomes clear when the limit is given its usual 'ε … δ' definition.

We may generalize this for a function between two topological spaces: a function $f: X \to Y$ is continuous if the pre-image of every open set of Y is open in X. We may think of this as a kind of 'supervaluating' over all metrics, preserving what is structurally common to all continuous functions, whatever the underlying metric. This may go some way to allaying the concern, raised in section 3, that the choice of any particular metric is arbitrary.[6] Weber and Colyvan (2010) give a topological version of the sorites paradox. Kelly (1996) appeals to the topological definition of continuity to characterize empirical methods, regarding possible data streams as infinite sequences of discrete inputs, and open sets as propositions that are empirically verifiable by finite sequences of such inputs.

We philosophers can outsource a lot of our problems, and let practitioners of other fields do a lot of the hard work for us. In this case, various beautiful theorems involving continuity that mathematicians have proven can do philosophical work. Start with the *intermediate value theorem:*

> *If f is a real-valued continuous function on the interval [a, b], and u is a number between f(a) and f(b), then there is a c ∈ [a, b] such that f(c) = u.*

The underlying idea is simple. Think of a continuous function as one that you can trace without your pen ever leaving the page. Start at its value at the beginning of a closed interval, and trace it until you reach its value at the end. Along the way you must have crossed every intermediate value between those two endpoint values—in particular, any designated intermediate value.

[6] Thanks to Mark Colyvan here.

We can use the intermediate value theorem to argue that at any time there must be a pair of antipodal points on the Earth that have the same temperature (and pressure too!) More philosophically, Joyce (1998) appeals to this theorem in his argument for probabilism, the thesis that rational credences obey probability theory. He shows that if your credences violate probability theory, then they are *accuracy-dominated*: there exists a probability function that is closer to the truth in every possible world. The intermediate value theorem supports a key step in the argument.

Or consider Brouwer's fixed point theorem:

> *Any continuous function f from a closed interval of the real line to itself has a fixed point—a point x_0 such that*
>
> $$f(x_0) = x_0$$

Arntzenius and Maudlin (2010) ingeniously invoke this theorem to resolve a paradox concerning time travel. An old-fashioned camera takes a picture of a developed black-and-white film—a 'negative'—that leaves a time machine. The picture is developed, and the negative put in the time machine, sent back to the time at which the picture is to be taken, and leaves the time machine then. The trouble is that a negative has the complementary shades to the object of which it is a picture—a dark grey object has a light grey negative, and so on—so the story appears to be contradictory. But there is a neat solution. Represent the mapping from a shade of grey to its complementary shade as a continuous function on [0, 1], with 0 representing pure black and 1 representing pure white. By Brouwer's theorem, it has a fixed point: a shade of grey that is its own complement. So if the developed film is uniformly that shade, the story is consistent!

I adverted to a famous theorem involving continuity in my argument that there can be motion without velocity: the existence of an everywhere-continuous function that is nowhere-differentiable. Interestingly, the Weierstrass function is the limit of an infinite sequence of functions, each of which is *everywhere*-differentiable. The anomalous behaviour of the function kicks in suddenly 'at infinity', in this sense behaving nothing like the functions that approach it.

3.3 Discontinuity at Infinity

This is an instance of another philosophically important phenomenon, which we might call *discontinuity at infinity*. This is not the place to characterize this technical notion rigourously, but roughly it involves the failure of a natural extension to the definition of continuity to cases where we can make sense of a function's behaviour at infinity:

$$lim_{x_n \to \infty} f(x_n) = f(\infty)$$

where $<x_n>$ is an increasing sequence. It could be a sequence of ordered *functions*, with *f* representing a binary property that another function could have or not. For example, everywhere-differentiability of a function could be represented by a 1, its failure represented

by a 0; the Weierstrass function scores a 0, even though it is the limit of a sequence of functions each of which scores a 1.

Again, we recognize discontinuity at infinity in various philosophical examples. It figures in certain paradoxical decision problems that involve infinity in some way. Nover and Hájek's (2004) *Pasadena game* has a pathological (indeed, undefined) expected utility that is the limit of a sequence of perfectly well-behaved expected utilities. Arntzenius, Elga, and Hawthorne's (2004) 'Satan's Apple' involves an infinite sequential choice problem that becomes paradoxical only when the limit at infinity is reached. And Pollock's problem of when to open the Ever-better wine displays a similar discontinuity at infinity: the option of waiting forever and never opening the wine is discontinuously worse than all of the options that approach it. Bartha, Barker, and Hájek (2013) characterize in more detail decision problems that are discontinuous at infinity, and discuss some methods for approaching them.

While I applaud using mathematical theorems to support philosophical points (and I have just done so myself), we should keep in mind Benacerraf's (1967) caution about deriving philosophical conclusions from mathematical theorems: *When somebody purports to establish a philosophical thesis with a mathematical theorem, they must have assumed some philosophical premises.* In fact, keeping this caution in mind is a good philosophical heuristic in its own right! Examples of apparent violations of Benacerraf's cautionary words include:

- Putnam (1980): the Löwenheim-Skolem theorem shows that realism is false.
- Lucas (1961) and Penrose (1989): Gödel's theorem shows that minds are not machines.
- Various people: the Dutch book theorem shows that credences are rationally required to be probabilities.

None of these theorems establish what they are purported to without the aid of ancillary philosophical premises. Spelling them out clarifies what the associated arguments really should be, and in doing so fixes targets for further debate. I, for example, assumed that a particle could in principle move according to a Weierstrass function.

3.4 Continuity Reasoning in Philosophical Methodology

But let's continue with continuity reasoning, this time applied at a 'meta'-level. Continuity considerations can be applied to philosophical methodology itself. Consider Priest's (1994) *Principle of Uniform Solution:*

> *If two paradoxes are of the same kind, then they should have the same kind of solution.*

Think of paradoxes as situated in 'paradox space', and their solutions as situated in 'solution space'. The Principle can be regarded as a requirement that nearby paradoxes should be mapped to nearby solutions.

Again, we may ask: what is the appropriate metric—when are two paradoxes of the same kind? Smith (2000) presses this concern, arguing that two paradoxes may be similar at one level of abstraction, while being dissimilar at another. And is discontinuity sometimes acceptable—two paradoxes of the same kind having quite different solutions? I suggest

that it is. Think of a limiting case: the *very same* paradox may get two quite different solutions, both adequate. (They may target different premises, for starters.) All the more, then, it would seem that two nearby paradoxes could have two quite different solutions, both adequate. Note that the argument that I have just given itself invokes a kind of continuity reasoning!

Notice how similar the Principle of Uniform Solution is to van Fraassen's (1989, p. 236) 'Symmetry Requirement':

> *Problems which are essentially the same must receive essentially the same solution.*

More generally, I think that the symmetry requirement is kindred with continuity reasoning. Both are useful heuristics. Jaynes (2003), for instance, appeals to the symmetry requirement in his defence of the principle of indifference, applying the principle to transformations of a probability problem that leave it essentially unchanged.

3.5 Continuity and Modal Induction

Earlier I emphasized the important role that continuity reasoning can play in induction, ampliatively inferring a conclusion about the actual world from a premise about this world. Continuity reasoning can also play an important role in what we might call *modal induction*: ampliatively inferring a conclusion about the space of *possible worlds* from a premise about this space.

In my (2014) I offer various methods for showing that something, call it *X*, is possible. One such method can be regarded as employing continuity reasoning. It follows this schema:

1) *Almost-X* is possible.
2) The small difference between *almost-X* and *X* makes no difference to possibility.

Conclusion: *X* is possible.

Chalmers (1996) argues along these lines that physical-duplicate zombies are possible (beings that have no conscious experiences, but that are physically identical to normal agents that have conscious experiences). Functional-duplicate zombies (functionally identical to normal agents), he argues firstly, are possible; and moving from functional-duplicates to physical-duplicates will not make a difference to what's possible.

4. Mismatch of Degrees

It's a problem for a putative analysis of some concept, *C*, if *C* comes in degrees that vary continuously, while the analysans has discontinuous 'jumps', or vice versa. More generally, it's a problem if there is a mismatch of the degrees of the analysandum and the analysans. A notable case of this is when one side of a putative analysis comes in degrees, while the

other does not. This in turn may involve a mismatch of *vagueness*, one side admitting borderline cases, the other not.

4.1 The Analysandum Does Not Come in Degrees, the Analysans Does

Berkeley was fond of saying that *to exist is to be perceived*. (For some reason, he was particularly fond of saying it in Latin.) Now, existence does not come in degrees—it is the ultimate on/off property or attribute. But offhand, being perceived does. Think of perceiving a table in a room that is initially totally dark, slowly turning the lights up using a 'dimmer' dial. Or think of things on the periphery of your visual field. Moreover, these provide borderline cases of being perceived; but it is hard to make sense of borderline cases of existence.

To be sure, we might reply on behalf of Berkeley that 'being perceived' is a threshold notion—for example, perceiving the table suddenly begins at a certain minimal light level. But that introduces a disquieting discontinuity, and an apparent arbitrariness in the setting of the threshold. Then again, we might use one of the many responses to arbitrariness that I detailed in section 2.1, so I do not claim this objection is decisive. Berkeley himself replied that everything that exists is perceived by God; and presumably he would add that God's perceptions don't come in degrees.

Personal identity seems to be 'all or nothing', yet various analyses of it involve things that come in degrees (e.g. psychological or bodily continuity—not the same sense of 'continuity' as before!). But—another useful philosopher's technique—one person's modus ponens is another's modus tollens. And Parfit (1984) tollenses where others ponens, arguing that personal identity indeed comes in degrees. (If it does, my earlier insouciance about the discontinuity of the *Alan Hájek* function may need revisiting.)

Now turn things around (another philosophical heuristic in its own right!).

4.2 The Analysandum Comes in Degrees, the Analysans Does Not[7]

Hume (1748/1902) defines a miracle as a violation of a law of nature. But arguably, being a miracle comes in degrees, whereas being a law of nature does not, and thus violating a law of nature does not. Some miracles are more *miraculous* than others. Arguably, a resurrection would genuinely violate a law of nature. (If not, then Hume's definition is in even more trouble, since this about as miraculous an event as any attested to.) But other miracles need

[7] If you or your students have trouble remembering which is which, here is a mnemonic: the analysan*dum* is the *dumb* thing (the concept as it is found in the wild, pre-reflection), and the analys*ans* is the *ans*wer (as provided by some thoughtful philosopher). I gather that others have hit upon this mnemonic before me, including David Sosa.

not. For example, the Red Sea parting on Moses' command need not. Consistent with the laws, one water molecule after another could spontaneously move this way or that way, collectively constituting the parting of the sea. To be sure, the event is extraordinarily *improbable*, given the laws. Perhaps, then, a miracle should somehow be defined in terms of improbability instead. After all, improbability comes in degrees, and is thus fit to match the degrees of the analysandum.

According to the von Mises (1957)/Church (1940) analysis of randomness, a sequence is random iff *every* recursively specified subsequence has the same relative frequency of every attribute. If this analysis is unfamiliar to you, and indeed if even the terminology in the analysans is unfamiliar to you, so much the better—the power of this heuristic will be revealed all the more. For even if the right-hand-side makes you glaze over, you know that *universal quantification does not come in degrees*. (The same is true of existential quantification.) Either it is (entirely) true that *every* blah-blah has yadda-yadda, or it is (entirely) false; the analysans does not come in degrees. But surely randomness comes in degrees. Suppose we toss a coin indefinitely. Consider interspersing the completely random sequence (say)

H H T H T T T T H...

with the obviously *non*-random alternating sequence

H T H T H T H T...

to yield:

H H H T T H H T T H T T T H T T H...

Surely this resulting sequence is *partially* random. But without paying any attention to the details of the von Mises/Church analysis, you know that it cannot deliver that verdict. (In fact, its verdict is that the sequence is *non*-random—entirely so!)

This account of randomness was intended to undergird von Mises' *frequentist* account of (objective) probabilities: they are relative frequencies in random sequences of trials. Philosophers now mostly agree that frequentism provides a bad analysis of probability (see my 1996 and 2009 for many arguments for this conclusion). But *a bad analysis can provide a good heuristic*. It will typically not be a total failure—it will typically get a wide range of central cases right. (If it did not, it presumably would never see the light of day.) While failing to give necessary and sufficient conditions for its analysandum, it may nevertheless usually be a reliable guide to the analysandum, reliable enough to serve as a useful way to think about the associated concept.

This brings us to the next heuristic—or better, set of heuristics.

5. Replace Non-Extensional Notions with Extensional Surrogates

Kahneman and Tversky (1982) have famously contended that we are bad at reasoning probabilistically—witness our tendency to neglect base rates or to commit the conjunction

fallacy on various probability questions.[8] For example, people are prone to think that if they test positive for a disease when given a test that is 95% reliable—it has a 5% chance of giving false positives and a 5% chance of giving false negatives—the chance that they have the disease is 95%. This neglects the prior probability ('base rate') that they have the disease in the first place, which may render that chance much smaller (or indeed, higher). The importance of this prior probability is obvious when we consider an extreme case, in which a given person has no chance of having the disease; then, a positive test result is guaranteed to be a false positive. (By analogy, if a man gets a positive result on a pregnancy test, it's obvious what he should conclude!) By the way, witness the heuristic power of considering extreme cases, which I emphasize in my (2014).

Gigerenzer (1991) has found that if probability problems are rephrased in terms of frequencies, we fare much better. This is not to say that probabilities *are* frequencies (they are not): just that frequentist thinking is a good heuristic for probabilistic thinking. So, translate a probability problem into a parallel frequency problem, or if you like, *pretend* that probabilities are frequencies. Imagine, say, that you are one of 10,000 people who have taken the test, and that it showed positive. Now suppose you learn that the base rate of the disease is 1%, so only 100 people in the population have the disease, and 9,900 don't (according to the pretence); 95 of the former group (truly) showed positive, and 495 of the latter group (falsely) showed positive. That is, less than 1/5 of the people in your shoes actually have the disease. That should be comforting. Frequentist thinking makes the point intuitive and vivid.

Inspired by this, I propose a related heuristic: *replace intensional notions with extensional surrogates*. An intensional notion is one for which truth value may fail to be preserved under replacement of co-referential expressions; an extensional notion is one for which truth value is preserved.[9] I submit that extensional notions are easier to think about and to deal with; intensional notions are more *opaque* to us. The typical cases that I will consider are ones in which we begin with a *modal* notion, and replace it with some *quantificational* notion. But the heuristic covers other cases as well.

In offering this heuristic I am not suggesting that a given extensional *replaceans* needs to provide an analysis or reduction of the original intensional *replaceandum*. The heuristic is entirely analytically and metaphysically innocent. I am suggesting only that there is heuristic value in *reasoning* with the extensional surrogate; it is a proposal for guiding the mind. To be sure, one hopes that the surrogate resembles the original closely enough that the mind does not go badly wrong when so guided, at least in typical or important cases. If the surrogate does provide a genuine analysis or reduction (in which case calling it a 'surrogate' undersells it), so much the better.

Some examples are well entrenched in philosophical thinking:

- Replace talk of *necessity* or *possibility* with talk of *what's true at all* or *some (accessible) possible worlds*.
- Replace talk of *counterfactuals* with talk of *what's true at the nearest antecedent worlds*.

[8] This section benefited from discussion with Sharon Berry, Yooav Isaacs, Hanti Lin, Daniel Nolan, and Wolfgang Schwarz.

[9] I follow the standard definition here, although it will not be important to adhere to it strictly. (That would hardly be in the spirit of my project!) In fact, soon I will loosen up the heuristic.

Here at least, most philosophers think that the extensional surrogates really are equivalent to the original intensional notions. In some other cases, the alleged equivalence is controversial, but the heuristic value is there nonetheless. For example:

- Replace talk of *indeterminacy* with talk of *what's true on some but not all admissible precisifications*, and replace talk of *what's determinately true/false* with talk of *what's true/false in all admissible precisifications*.
- Replace *indicative conditionals* with *material conditionals*. (Material conditionals are extensional since they are truth functional, and the extension of a sentence is its truth value.)

Or perhaps better is a two-step replacement:

- Replace *indicative conditionals* with *strict conditionals*, and replace *them* with *material conditionals true in all possible worlds*.

The first step replaces an intensional notion with another intensional notion, but one that has a simple extensional translation.

Then there are analyses whose failure is relatively uncontroversial nowadays, yet still they may be useful ways to think about the relevant concepts, if only as an act of pretence:

- Replace talk of *laws of nature* with talk of *universal generalizations over individual entities*.
- Replace talk of *causation* with talk of *constant conjunction of event-types*.

Indeed, a large part of the failure in each case stems from the extensional analysans not capturing the *intensionality* of the analysandum![10]

I have so far discussed the handling of intensional notions. But even more recalcitrant to our natural modes of thinking are *hyperintensional* notions—ones for which truth value may fail to be preserved even under replacement of *necessarily* co-referential expressions. Properties, at least when finely individuated, are often taken to be examples—think of *triangular* and *trilateral* as distinct properties, yet necessarily co-present or co-absent. Still, there is an extensional surrogate:

- Replace talk of *properties* with talk of *sets of possible individuals*.

Hyperintensional notions count as intensional according to the definition that I gave; but it is common to give special status to them, distinguishing them from what we might call the

[10] A reason why these extensional analyses clearly fail, where some considered earlier appear to succeed, may be that these quantify over extensional entities, whereas the successful ones quantified over entities that themselves might be regarded as intensional (possible worlds, a modal notion). It is less clear that the analysis of (in)determinacy is successful; but then, perhaps it is less clear whether the notion of a precisification being *admissible* is itself intensional. (If it is tacitly modal, then presumably it is.) To the extent that an analysis traffics in entities that are themselves intensional, perhaps it is not truly an extensional analysis after all. I bracket that concern here, trusting that applications of my heuristic are easy to recognize in any case.

merely intensional. To avoid any terminological confusion, we might do better to restate our heuristic: *replace non-extensional notions with extensional surrogates*. This covers both ways for a notion to fail to be extensional: (merely) intensional, and hyperintensional. Set-theoretic treatments of concepts are extensional, and they often provide helpful replacements of both kinds of concepts.

And so it goes. We might come up with our own ways of unpacking a non-extensional concept in an extensional way. I like this one:

- Replace talk of *norms* with *universal quantification over all norm-abiding agents*. For example, replace 'one has a moral obligation to treat others as ends rather than means' with 'all morality-abiding agents are agents who treat others as ends rather than means'.

Replace non-extensional notions with extensional surrogates is a special case of a more general heuristic: *replace hard notions with easier surrogates*. That's rather vague, but the idea is to replace concepts that confound us with close relatives that we understand better, that come more naturally to us—for example, that are expressible using first-order logic and set theory. We must be careful that there is not too much distance between the original concepts and their surrogates, for if there is, we run the risk that the surrogates will *mis*guide us. In an especially happy replacement, the replaceans has the same logic as the replaceandum, sanctioning exactly the same inferences. This is clearly the case when replacing probabilities with frequencies—the latter obey the probability axioms (with finite additivity). It is arguably also the case with the extensional replacements of necessity and possibility (with suitable choices of accessibility relations), and of counterfactuals. But even when it is clearly not the case, the replaceans can serve as a guide to reasoning—defeasible, to be sure. Mill's methods for identifying causation are really methods for identifying constant conjunction; and while causal inference has come a long way since Mill, they still provide good rules of thumb. We should just remember that good rules of thumb are not the last word.

Replace non-extensional notions with extensional surrogates is a powerful heuristic, I think, because it is easier to *picture* things when they're extensional. This brings me to my final heuristic.

6. Draw a Picture

This heuristic again draws on a psychological fact about us, this time a familiar one. Conceptual relationships that may be obscure when cloaked in words or symbols often leap out at us when we represent them visually. Think of the power of Venn diagrams to represent set-theoretic relationships. And the surrogates above that can be cast set-theoretically immediately lend themselves to Venn diagrams. The foregoing heuristic, then, works particularly well in tandem with this one: *translate from non-extensional to extensional notions, then diagram the latter.*

Example. If you draw a Venn diagram of the disease test example above, with areas representing classes of individuals roughly proportional to their respective population sizes, the importance of the base rate becomes even more obvious.

Example. Thanks to the Kripke semantics for various systems of modal logic, we can easily diagram how different modal systems, corresponding to different assumptions on an 'accessibility' relation among worlds, validate different inferences.

Example. Having given his own possible worlds analysis of counterfactuals, Lewis (1973a) goes on to present various counterfactual fallacies, such as antecedent-strengthening:

$$p \mathrel{\square\!\!\rightarrow} q$$
$$\therefore (p \,\&\, r) \mathrel{\square\!\!\rightarrow} q.$$

('$\square\!\!\rightarrow$' symbolizes the counterfactual conditional.) It may not be immediately obvious that this is not a valid argument form. But he goes on to draw diagrams that allow one to *see* easily that it is not.

However, the *draw a picture* heuristic can stand alone, unaided by the previous heuristic. For example, it helps to diagram causal relationships even without entertaining the fiction that causation is merely constant conjunction—either with Lewis-style (1973b) 'neuron diagrams', or Pearl-style (2009) causal networks.

Physicists know well the power of pictures—for example, drawing a Minkowski space–time diagram rather than performing a deduction using the mathematics of special relativity. When a rod with clocks at each end moves relative to you, the clock at the front runs behind the clock at the back. Showing this is relatively easy with a diagram, relatively hard with the mathematics. Mathematicians are similarly well versed in the heuristic utility of pictorial representations of abstract relationships. Indeed, Nelson (1993) is an entire book of proofs without words or symbols. Again, we philosophers can import some of our heuristics from other disciplines.

7. Some Metaphilosophical Ruminations

It is another psychological fact about us that we are limited in our ability to see what follows from what. (The Wason selection task famously brings this out.) We are prone to drawing illicit inferences from our premises, and we are often blind to their consequences, which are sometimes unwelcome. Here, logic is our great tool, our consistency and validity policeman. But I must emphasize again that it is by no means our only such tool.

I can characterize a good chunk of the philosophy that I find congenial with the slogan *'making our implicit commitments explicit'*. We collect, deploy and systematize intuitions, analyze arguments, conduct thought experiments, refine definitions, adduce normative constraints, and so on; and when we are doing our job properly, we check for the consistency and for any unwelcome consequences of our products (and celebrate the welcome ones). The heuristics are to a large extent aids to these enterprises.

Of course, you don't need to be a philosopher to make our implicit commitments explicit. Mathematicians do it too, and they remind us what a worthy enterprise it can be. If my slogan sounds like it trivializes philosophy, we should remember that exemplifying it

may be no easy feat. Nobody thinks that it's easy to make explicit how our commitment to certain basic facts about positive integers, addition and exponentiation implicitly commit us to Fermat's Last Theorem.

If we think of logic as an all-purpose tool for checking what follows from what, and what is compatible with what, at a high level of abstraction, then the heuristics collectively form more of a Swiss army knife. Some of them have rather more specific targets or uses. 'See definite descriptions in neon lights' is a bit like a Swiss army knife's scissors: useless for many occasions, but exactly what you need for some. We want heuristics that have some generality, but that also have real bite. Too general, and they become useless: for example, 'make an insightful point!' Too specific, and they become next to useless again: for example, 'when someone argues that space is purely relational, reply with Kant's problem of incongruous counterparts!' The best heuristics lie somewhere in the middle of the spectrum: general enough to be applicable in a wide range of cases, but not so general as to be empty. But of course the heuristics can vary considerably in their generality among themselves, and still earn their keep. The Swiss army knife's scissors may have a greater range of uses than its corkscrew, but on certain occasions the latter is exactly what you need.[11]

8. Conclusion

I've given a small sampler of some philosophical heuristics. I have chosen them because there are some nice interconnections among them, and I have tried to present some of them in considerable detail to illustrate their fruitfulness from a number of angles. But I could have chosen any number of others.

As well as studying the *content* of our best scientific theories, we study the *methodology* of the best scientists. (See, for instance, Harper's recent (2012) book on Newton's methodology.) Doing so not only tells us something about how science gets done; it can also inspire us as to how to do it well. Similarly, studying the methodology of the best philosophers not only tells us something about how philosophy gets done; it can also inspire us as to how to do it well. I have learned from them many of the philosophical heuristics on my ever-growing list.

To the extent that we fail to pay attention to such heuristics, we miss out on rich resources for philosophical thinking. Yet I think that by and large, philosophers have been singularly unreflective about them. I invite you to reflect on them with me.

References

Arntzenius, Frank, Adam Elga, and John Hawthorne (2004): 'Bayesianism, Infinite Decisions, and Binding', *Mind*, 113: 251–83.

Arntzenius, Frank and Maudlin, Tim (2010): "Time Travel and Modern Physics", in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Spring 2010 Edition)*. URL = <http://

[11] Now that you have reached the conclusion of this long paper, perhaps this is one of them.

plato.stanford.edu/archives/spr2010/entries/time-travel-phys/>. Accessed September 21, 2015.
Bartha, Paul, John Barker, and Alan Hájek (2013): "Satan, Saint Peter, and Saint Petersburg: Decision Theory and Discontinuity at Infinity", *Synthese*, 191(4): 629–60.
Benacerraf, Paul (1967): "God, the Devil, and Gödel", *The Monist*, 51: 9–32.
Carnap, Rudolf (1952): *The Continuum of Inductive Methods*, Chicago: University of Chicago Press.
Chalmers, David J. (1996): *The Conscious Mind*, Oxford: Oxford University Press.
Church, A., 1940: "On the Concept of a Random Sequence", *Bulletin of the American Mathematical Society*, 46: 130–5.
Elga, Adam (2004): "Defeating Dr. Evil with Self-Locating Belief", *Philosophy and Phenomenological Research*, 69: 383–96.
Gigerenzer, Gerd (1991): "How to Make Cognitive Illusions Disappear: Beyond 'Heuristics and Biases'", *European Review of Social Psychology*, 2: 83–115.
Hájek, Alan (1996): "'*Mises Redux*'—*Redux*: Fifteen Arguments Against Finite Frequentism", *Erkenntnis*, 45: 209–27. Reprinted in *Philosophy of Probability: Contemporary Readings*, ed. Antony Eagle, Routledge 2010.
Hájek, Alan (2009): "Fifteen Arguments Against Hypothetical Frequentism", *Erkenntnis*, 70: 211–35.
Hájek, Alan (2014): "Philosophical Heuristics and Philosophical Creativity", in Elliot Samuel Paul and Scott Barry Kaufman (eds.) *The Philosophy of Creativity*. Oxford: Oxford University Press, 288–317.
Harper, William L. (2012): *Isaac Newton's Scientific Method*, Oxford: Oxford University Press.
Hume, David (1748/1902), *Enquiries Concerning the Human Understanding and Concerning the Principles of Morals*, ed. L. A. Selby-Bigge, Oxford: Clarendon Press, Second Edition, 1902.
Jaynes, E. T. (2003): *Probability Theory: The Logic of Science*. Cambridge University Press.
Joyce, James M. (1998): "A Non-Pragmatic Vindication of Probabilism", *Philosophy of Science*, 65: 575–603.
Kelly, Kevin (1996): *The Logic of Reliable Inquiry*, Oxford: Oxford University Press.
Lewis, David (1969): *Convention*, Cambridge: Harvard University Press.
Lewis, David (1973a): *Counterfactuals*, Oxford: Blackwell.
Lewis, David (1973b): "Causation", *Journal of Philosophy*, 70: 556–67.
Lewis, David (1994): "Humean Supervenience Debugged", *Mind*, 103: 473–90.
Lucas, J. R. (1961): "Minds, Machines and Gödel," *Philosophy*, 36: 112–27.
Lycan, William (2009): "Giving Dualism Its Due", *Australasian Journal of Philosophy*, 87 (4): 551–63.
Nelson, Roger B. (1993): *Proofs Without Words*, The Mathematical Association of America.
Nover, Harris and Alan Hájek (2004): "Vexing Expectations", *Mind*, 113: 305–17.
Nozick, Robert (1993): *The Nature of Rationality*, Princeton: Princeton University Press.
Parfit, Derek (1984): *Reasons and Persons*, Oxford: Oxford University Press.
Pearl, Judea (2009): *Causality*, Cambridge: Cambridge University Press, 2nd edition.
Penrose, R. (1989): *The Emperor's New Mind*, Oxford: Oxford University Press.
Pollock, John (1983): "How Do You Maximize Expectation Value?", *Noûs*, 17: 409–21.
Pólya G. (1957): *How to Solve It*, 2nd ed., Princeton: Princeton University Press.
Priest, Graham (1994): "The Structure of the Paradoxes of Self-Reference", *Mind*, 103: 25–34.
Putnam, Hilary (1980): "Models and Reality" *The Journal of Symbolic Logic*, 45 (3): 464–82.

Sider, Ted (2002): "Hell and Vagueness", *Faith and Philosophy*, 19: 58–68.
Smith, Nicholas J. J. (2000): "The Principle of Uniform Solution (of the Paradoxes of Self-Reference)", *Mind*, 109: 117–22
Stalnaker, Robert (1968): "A Theory of Conditionals", in N. Rescher (ed.), *Studies in Logical Theory*, Oxford: Oxford University Press, 98–112.
Stalnaker, Robert (1981): "A Defense of Conditional Excluded Middle", in Harper, W. L., Stalnaker, R., and Pearce, G. (eds.), *Ifs*, Reidel, Dordrecht, 87–104.
Tversky, Amos and Daniel Kahneman (1982): "Judgment Under Uncertainty: Heuristics and Biases", in Daniel Kahneman, Paul Slovic and Amos Tversky (eds.), *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press, 3–22.
Unger, Peter 1979, 'There Are No Ordinary Things', *Synthese*, 41: 117–54.
van Fraassen, Bas (1989): *Laws and Symmetry*, Oxford: Oxford University Press.
van Inwagen, Peter (1990): *Material Beings*, Ithaca: Cornell.
von Mises R., 1957, *Probability, Statistics and Truth*, revised English edition, New York: Macmillan.
Weber, Zach and Mark Colyvan (2010): "A Topological Sorites", *The Journal of Philosophy*, 107 (6) (June): 311–25.
Williamson, Timothy (1994): *Vagueness*, London: Routledge.

CHAPTER 20

DISAGREEMENT IN PHILOSOPHY

Its Epistemic Significance

THOMAS KELLY

1. Introduction

Despite occasional declarations that philosophy is now in a position to move beyond the contentiousness of its past, pervasive and intractable disagreement seems to be among its characteristic features. In any case, such disagreement is a notable feature of philosophy's present as well as its history, and shows no signs of abating. The fact that philosophers disagree as often as they do, and in the ways that they do, is itself a phenomenon that gives rise to philosophical questions. Some of these questions are epistemological. For example, how (if at all) should one revise one's philosophical views about controversial issues, given one's awareness that those views are not shared (and indeed, are explicitly rejected) by apparently competent philosophers? Should we refrain from holding confident philosophical beliefs when informed philosophical opinion is divided about their truth? Does the lack of consensus in philosophy mean that *philosophical knowledge* is off the table, at least in areas where it is unrealistic to think that consensus is in the offing?

This chapter explores the epistemological significance of disagreement in philosophy in the light of some currently prominent theories of disagreement. I will be particularly concerned with whether the kind of controversy that we find in philosophy warrants a certain kind of skepticism about the subject. Call someone a *philosophical agnostic* just in case he or she refrains from holding confident opinions about philosophical questions: either he or she suspends judgment about philosophical questions, or at best, holds relatively weak, diffident opinions about their correct answers.[1] A natural thought that has occurred to a

[1] Care is needed here to ensure that the stance of the philosophical agnostic ends up being coherent. For example, it would not do to identify *suspending judgment about whether p* with *investing credence .5 in p*: given that some philosophical theses are logically stronger than others, giving .5 credence to every philosophical thesis would quickly result in incoherence. I will assume for the sake of argument there is some way of representing the stance of the philosophical agnostic so that it is coherent.

number of philosophers is this: given the kind of disagreement that we find in philosophy, we really ought to be philosophical agnostics—at least when it comes to the many philosophical issues that are controversial among well-informed, competent members of the professional community.[2] To the extent that one does hold confident opinions about contested philosophical matters, those opinions are less than fully reasonable. (At least, assuming that one is aware of their contested status.) Is this natural thought correct?

Ideally, the best way to answer this question would be the following: take the correct general account of how we should respond to disagreement and apply it to the philosophical case by plugging in various empirical facts about the disagreements that we find in philosophy. Unfortunately, there is no generally agreed upon account of how we should revise our beliefs in the light of disagreement: the epistemology of disagreement is itself an area about which informed philosophical opinion is deeply divided.[3] This might seem to leave us in the following situation: whether one sees the lack of agreement within philosophy as a compelling reason for philosophical agnosticism will depend upon which general view about the epistemology of disagreement one accepts. If, for example, one accepts a view on which it can be reasonable to steadfastly maintain one's opinions even when they are contested by those whose judgment one has no independent reason to discount, then one will not see the lack of agreement within philosophy as warranting agnosticism. On the other hand, if one adopts a more conciliatory view about the epistemology of disagreement, one will see the profound lack of agreement as a compelling reason to adopt the stance of the agnostic. If this were right, then whether philosophical agnosticism is ultimately compelling would depend (naturally enough) on which of the various rival views about the epistemology of disagreement ultimately turns out to be correct.

Although natural, I think that the line of thought sketched in the last paragraph is mistaken. My contrary thesis is this: there is *no* plausible view about the epistemology of disagreement on which philosophical agnosticism is compelling.

In order to explore this issue, I will survey some prominent alternative views about the epistemology of disagreement and examine what happens when they are applied to the special case of philosophy. I will begin with views about disagreement on which the case for philosophical agnosticism seems weakest, and then proceed to consider, in increasingly greater detail, views on which the case seems strongest, at least prima facie. (As this suggests, the amount of space and attention devoted to a given view will *not* reflect my own estimate of the chance that it is correct, but rather the plausibility with which it might be thought to deliver the agnostic's conclusion.) I will suggest that even when the latter views are assumed to be true, there is no clear path from the kind of disagreement we find in philosophy to the skeptical attitude about the subject that is characteristic of the philosophical agnostic.

[2] Previously published work that engages with this thought to at least some degree, and with varying levels of sympathy, includes Armstrong (2006), Frances (2010, 2013), Fumerton (2010), Goldberg (2009, 2013), Kornblith (2010), and van Inwagen (1996). I regret that the strict word limits of the present volume prevent me from engaging with this literature in the kind of detail that it deserves, although much that I say here bears on claims made by these authors.

[3] For a sampling of this, see the essays collected in Feldman and Warfield (2010) and Christensen and Lackey (2013).

Before beginning the survey, I should emphasize that my concern is not with philosophical agnosticism in general, or with the variety of grounds that might be offered in favor of such a stance, but rather with the claim that we should be philosophical agnostics *because of something having to do with disagreement*, or with the distribution of opinion that we find among philosophers. Perhaps we should be philosophical agnostics for reasons unrelated to the phenomenon of disagreement. For example, one might think that even the best philosophical arguments are simply not very strong compared to the kinds of arguments we find in other domains[4] and conclude on that basis that we should not hold confident philosophical opinions. Nothing that is said in what follows will address such possibilities. Indeed, because our exclusive concern here is with the potential skeptical import of philosophical disagreement, it will be helpful to proceed as follows: we assume, optimistically, that we are or would be in a position to attain philosophical knowledge, or at least justified philosophical beliefs, *before facts about philosophical disagreement or the distribution of opinion among philosophers are taken into account*.[5] As we will see, there are prominent views about the epistemology of disagreement according to which the right kind of disagreement can undermine the rationality of continuing to hold a given belief, even if one arrived at that belief in a rationally impeccable manner, and even if one's belief would have amounted to knowledge in the absence of the disagreement. Applied to the case of philosophy, the relevant thought would be the following: given the kinds of disagreements we find among philosophers, one is not in a position to know or justifiably believe controversial philosophical theses, even if one would have been in a position to know and justifiably believe those theses in the absence of controversy.

2. Steadfast Views

Contributors to the epistemology of disagreement literature sometimes draw a rough but serviceable distinction between *steadfast* and *conciliatory* views. For example, David Christensen opens his (2011) paper with the following:

> Subtleties aside, a look at the topography of the disagreement debate reveals a major divide separating positions that are generally hospitable to maintaining one's confidence in the face of disagreement, and positions that would mandate extensive revision to our opinions

[4] Consider, for example, the pessimistic assessment of van Inwagen (2006: 52):

> There are certainly successful arguments, both in everyday life and in the sciences. But are there any successful philosophical arguments? I know of none. (That is, I know of none for any *substantive* philosophical thesis.)

(Van Inwagen does not draw the further skeptical conclusion, that the putative lack of successful philosophical arguments warrants agnosticism.) For a critique of the pessimism, see McGrath and Kelly, "Are There Any Successful Philosophical Arguments?".

[5] Compare the way in which, in discussions of Humean skepticism about induction, we proceed on the assumption that we have a great deal of ordinary perceptual knowledge about the world, i.e. that more radical kinds of skepticism than the one under consideration are false.

> on many controversial matters. Let us call positions of the first sort "Steadfast" and positions of the second sort "Conciliatory". (p. 1)

In this section, I will examine the case for philosophical agnosticism in the light of steadfast views of disagreement; beginning with the next section, I will turn to conciliatory views.

One kind of steadfast view involves commitment to a *permissive* conception of rationality, according to which there is frequently a significant amount of slack between a body of evidence and what can be reasonably believed given that evidence. In particular, some hold that rationality is sufficiently permissive to allow for the following kind of case: given the available evidence, some individuals reasonably believe p while other individuals reasonably believe not-p, with each group aware that there is informed disagreement about whether p. The kind of view that I have in mind is endorsed by Gideon Rosen in the following passage:

> It should be obvious that reasonable people can disagree, even when confronted with a single body of evidence. When a jury or a court is divided in a difficult case, the mere fact of disagreement does not mean that someone is being unreasonable. Paleontologists disagree about what killed the dinosaurs. And while it is possible that most of the parties to this dispute are irrational, this need not be the case. To the contrary, it would appear to be a fact of epistemic life that a careful review of the evidence does not guarantee consensus, even among thoughtful and otherwise rational investigators.
>
> (2001: 71–2)

Here, Rosen suggests that paleontologists who accept rival theories about the demise of the dinosaurs on the basis of shared evidence might be fully reasonable in steadfastly maintaining their views, despite being aware that their views are not shared (and indeed, are explicitly rejected) by a significant number of their professional colleagues. But, one might think, what holds for paleontologists holds for philosophers as well. Thus, an epistemic permissivist might reject the claim that disagreement requires philosophical agnosticism on the grounds that many, most, or even all of the questions that are controversial among philosophers constitute *permissive cases*.[6]

Some steadfast views do not presuppose the truth of a permissive conception of rationality. For example, according to what is sometimes called The Right Reasons View (Kelly 2005, Titelbaum 2015, Weatherson 2013), one is justified in steadfastly maintaining one's belief that p in the face of disagreement, provided one's belief is a correct response to the available evidence and arguments that bear on whether p. For purposes of illustration, suppose that, given the current state of play in the philosophical literature, the case for incompatibilism really is stronger than the case for compatibilism, and that at least some incompatibilists hold their view because they have recognized that this is so. According to the Right Reasons View, incompatibilists in this position are fully justified in maintaining their belief, notwithstanding the fact that they are fully aware that there are significant numbers of philosophically sophisticated compatibilists who vigorously disagree with them, and who are unpersuaded by the very considerations that they find compelling. More generally, a proponent of the Right Reasons View holds that, once one has taken into account all of the considerations that proponents of a rival view have offered in favor of

[6] For a recent exchange on epistemic permissivism, see White (2013) and Kelly (2013b).

their view and against one's own, one need not give any *additional* weight to the fact they disagree. So long as one correctly responds to the first-order evidence and arguments, one is fully justified in maintaining one's belief, and the rationality of believing as one does is not susceptible to being undermined or defeated by higher-order considerations having to do with how others have responded to the evidence and arguments.

Clearly, if we assume that the Right Reasons View is true, then the mere fact that there is a great deal of disagreement within philosophy does not show that we should be philosophical agnostics, or that it would be unreasonable to hold strong opinions about contested philosophical issues. For in any case in which someone holds a philosophical view on the basis of what are *in fact* good reasons, they are fully justified in sticking to their guns, even if that view is vigorously contested by others.

A related but somewhat less radical proposal is the Total Evidence View (Kelly 2010, Setiya 2012: 20–3, Scanlon 2014: 80), which shares features with both the Right Reasons View and the more conciliatory views that we will consider in section 3. On the Total Evidence View, when one learns that a person whose judgment one has good reason to respect has arrived at an opinion that conflicts with one's own, one acquires an additional piece of evidence against one's view (even if one was already aware of all of the considerations on which the other person bases her opinion). Typically, then, it will be rational to temper one's confidence to at least some extent. However, it does not follow that one should give up one's original opinion. Whether it is reasonable to retain that opinion (albeit with somewhat diminished confidence) depends on what is now one's total evidence, a body of evidence which includes both the fact that the other person judged as they did as well as the original evidence on which one based one's own opinion. And crucially, the proponent of the Total Evidence View insists, there is no guarantee that this enlarged body of evidence will mandate an attitude of agnosticism or suspension of judgment towards the original target proposition.

To make things more concrete, suppose as before that as things stand the case for incompatiblism is stronger than the case for compatibilism. Recognizing this, you take up the position as your own. You then learn that I (someone whose philosophical judgment you generally hold in high regard) have arrived at the opposite view on the basis of my own assessment of the arguments. (To temporarily avoid complications which we will take up in later sections, let's assume, unrealistically, that at this point you know of *no one else's* opinion about the issue.) Suppose next that upon learning that I am a convinced compatibilist, you reduce the credence that you give to incompatibilism to some degree. However, even after you take my opinion into account, you still regard it as significantly more likely that incompatibilism is true, to the point that you can correctly be described as retaining your belief in incompatibilism. It's not that you take yourself to have some *independent reason* for discounting my opinion relative to yours, as would be the case if, for example, I had confided in you that I had reviewed the arguments in a cursory fashion, or was suffering from a debilitating headache at the time. Rather, the only reasons that you have for thinking that I am more likely to be wrong than you about this particular issue are the very reasons on which you base your belief that incompatibilism is true in the first place.

Is it reasonable for you to retain your belief in these circumstances? Yes, says the Right Reasons theorist: given that the available arguments were sufficient to justify your belief in incompatibilism before you learned what I think, the mere fact that I turn out to disagree

with you is insufficient to undermine the rationality of your belief. No, say many proponents of conciliatory views: given that you have no independent reason to think that you were the one who did a better job in responding to the arguments, you should treat our original opinions even-handedly and thus suspend judgment. In contrast, a proponent of the Total Evidence View will insist that we have not yet been provided with enough information about the fiction in order to know whether it's reasonable for you to retain your belief. What it is reasonable for you to believe about the issue depends on what is now your total evidence, and that body of evidence includes both the philosophical arguments themselves, as well as the fact that I judge that compatibilism is correct. On the Total Evidence View, the fact that I believe as I do is a piece of evidence against your view; your total evidence is thus less supportive of belief in incompatibilism than it was before you learned of my opinion, or than it would be if I had independently arrived at the same conclusion as you. Whether it is nevertheless still adequate to justify your continuing to believe that incompatibilism is true will depend on such features of the case as the substantiality and strength of the original arguments for incompatibilism, as well as the weight of your reasons for thinking that I'm likely to be a good judge about the question—features that have not been stipulated.

How does the disagreement-based case for philosophical agnosticism fare if we assume the truth of the Total Evidence View? The answer to this question is not immediately obvious.[7] On the one hand, the Total Evidence View shares with the Right Reasons View a commitment to the idea that, when one arrives at a position in virtue of a correct assessment of the available arguments, one's belief has a certain resistance to being rationally undermined by the fact that others arrived at different conclusions, even when those others are generally competent philosophers possessed of good judgment. On the other hand, the Total Evidence View shares with conciliatory views a commitment to the idea that when such a person does arrive at a different conclusion, this counts as a piece of evidence that one has made a mistake, and moreover, that the more competent philosophers there are who disagree with one, the stronger the evidence is that one has made a mistake.[8] Indeed, on what I regard as the best version of the Total Evidence View, there will be cases in which even if the first-order evidence and arguments strongly support a given belief, it would be unreasonable to continue to hold that belief when one learns that sufficient numbers of those who are in the best position to judge think otherwise. (That is, even relatively strong first-order evidence can be swamped when enough higher-order evidence pointing in the opposite direction is amassed.) And one might contend that such cases are something like the norm in philosophy, where there is typically no shortage of formidable dissenters from one's preferred views.

[7] The Total Evidence View is sometimes criticized on the grounds that it fails to offer determinate advice in certain (incompletely specified) cases for which rival views do offer determinate advice. However, the most charitable way to understand this is as an expression of skepticism, on the part of proponents of the view, about whether there are any true disagreement norms at the level of generality that some have sought. For skepticism about the existence of such norms, see Feldman (2009), Kelly (2013a), and especially, Lasonen-Aarnio (2013).

[8] At least, provided there is some measure of independence in the way that these philosophers arrived at their views.

At this point, one thing seems clear: to the extent that a plausible case for philosophical agnosticism might be made on the assumption that The Total Evidence View is true, this is because of features that it shares with more conciliatory views. In general, the Total Evidence View is more hospitable to sticking to one's guns in the face of controversy than conciliatory views are. Because of this, *if* there is a sound disagreement-based case for philosophical agnosticism on the assumption that the Total Evidence View is true, then there will also be a sound disagreement-based case for philosophical agnosticism on the assumption that some conciliatory view is true. On the other hand, if, as I will suggest, there is no sound disagreement-based case for philosophical agnosticism on the assumption that some conciliatory view is true, then there will be no such case on the assumption that the Total Evidence View is true, either. Let us turn then, to an examination of conciliatory views.

3. Conciliatory views

Proponents of conciliatory views[9] tend to see the phenomenon of disagreement as mandating relatively extensive revisions to our opinions about many controversial matters. The case for such revision is often taken to be especially compelling when those who disagree are known to share certain intellectual virtues and to have reviewed the same evidence and arguments in arriving at their opinions. Here, for example, is Richard Feldman:

> [C]onsider those cases in which the reasonable thing to think is that another person, every bit as sensible, serious, and careful as oneself, has reviewed the same information as oneself and has come to a contrary conclusion to one's own . . . An honest description of the situation acknowledges its symmetry. . . . In those cases, I think, the skeptical conclusion is the reasonable one: it is not the case that both points of view are reasonable, and it is not the case that one's own point of view is somehow privileged. Rather, suspension of judgment is called for.
>
> (2006: 235)

One motivation for conciliatory views is that they deliver what strike many as the intuitively compelling verdicts about certain concrete examples. Consider, for example, Christensen's "Mental Math Case" (2007). In the Mental Math Case, you and I independently calculate our shares of the dinner tab. (We've agreed to divide the check evenly among everyone who was at dinner.) We know, on the basis of substantial track record evidence, that we are more or less equally competent when it comes to performing this general kind of calculation (in our long history of dining together, we almost always come up with the same answer, but on those occasions when we've come up with different answers, each of us has turned out to be correct approximately half the time). On this occasion, you arrive at the number $43 while I arrive at the number $45. A widely shared intuition is that, upon discovering this, both of us (including the person who in fact reasoned correctly, assuming that one of us did) should become much less confident of his or her original answer, and

[9] For representative statements, see Christensen (2007, 2011), Elga (2007), Feldman (2006), Kornblith (2010), and Matheson (2009).

that indeed, each of us should divide our credences more or less evenly between the two rival answers.

Consider also the way in which one might try to motivate conciliatory views by appealing to analogies involving inanimate measuring devices.[10] For example, suppose that I form my beliefs about the ambient temperature in some room by consulting my thermometer. (To avoid complications, we can suppose that I have no other access to the temperature of the room. Perhaps I'm in an adjacent room, looking at my thermometer through a window, and I know that the temperature of the room that I'm in is not reliably correlated with the temperature of the room in which the thermometer is located.) I have no reason to think that anything is amiss with my thermometer, so the beliefs that I form in this way are perfectly reasonable. However, I then discover that the reading returned by my thermometer is inconsistent with the reading returned by your thermometer, which is also clearly visible in the adjacent room. Unless I have some special reason to trust my thermometer over yours, it seems as though I should be agnostic about whose thermometer is correct. Certainly, it wouldn't be defensible to favor the reading of my thermometer simply because that's what *my* thermometer says, or because that's what I justifiably believed before I learned about your thermometer. Moreover, the same seems to be true even if your thermometer really is the one that's malfunctioning on this occasion, and mine is functioning perfectly. So long as I have no independent evidence that that is what is taking place, the mere fact that my thermometer is the one that's functioning properly on this occasion doesn't justify favoring what it reports over what yours reports. But, one might think, what holds for thermometers holds for believers as well: when I find myself in a disagreement with someone else, then, in the absence of some independent reason for thinking that I'm the one who is correct, I should suspend judgment, even if my original opinion was fully reasonable before I became aware of the disagreement.

For our purposes, a notable feature of these examples is the way in which they suggest that even beliefs with seemingly impeccable rational credentials can be undermined when a certain kind of conflict emerges. Suppose that in the Mental Math Case, you were the one who arrived at the correct answer via an impeccable calculation. In that case, it is extremely natural to credit you with *knowing* the correct answer (and with being in a position to rationally invest a high degree of confidence in it) prior to learning that I arrived at a different answer. Similarly, if I arrived at my original belief about the temperature by relying on a thermometer that was in fact functioning perfectly, then it is natural to credit me with knowing the temperature prior to the discovery that your thermometer says something else. Indeed, given the relatively precarious nature of philosophical reasoning, it is natural to think that the epistemic standing of beliefs arrived at via flawless arithmetical reasoning or via reliance on accurate thermometers generally compares favorably with the epistemic standing of beliefs to the effect that a given philosophical position is correct, even in cases in which the latter beliefs are arrived at *via* a correct assessment of the philosophical state of play. But if knowledge that has been arrived at *via* flawless arithmetical reasoning or *via* reliance on accurate thermometers can nevertheless be undermined by the emergence of disagreement, surely the epistemic credentials of one's philosophical opinions can be

[10] For a discussion of the strategy, see Kelly (2014).

undermined by one's awareness that apparently competent philosophers have arrived at incompatible views.

It might seem then, that adopting a conciliatory view about disagreement will generate strong and perhaps irresistible pressures towards philosophical agnosticism. However, I think that the path from conciliatory views to philosophical agnosticism is much more precarious than is often assumed. Indeed, closer inspection casts doubt on whether there is *any* plausible conciliatory view on which philosophical agnosticism is compelling.

As a first step towards seeing this, let's note that at least *some* prominent conciliatory views defended in the literature seem to allow for the holding of strong opinions about contested philosophical matters. Consider, for example, Adam Elga's "Equal Weight View" (2007). For Elga, a particularly threatening kind of disagreement occurs when one finds that one disagrees with *someone whom one regards as an epistemic peer.* In Elga's lingo, you count someone as an *epistemic peer* with respect to a given issue just in case: "… you think, conditional on a disagreement arising, the two of you are equally likely to be mistaken" (487). Christensen's Mental Math Case exemplifies a case of peer disagreement in this sense, or at least, is naturally read as such: given that you and I acknowledge that we are equally good at arithmetic, it is natural to think that, if we were asked prior to performing the calculation who was more likely to be wrong in the event of a disagreement, we would admit that we were equally likely to be wrong; in this respect, we regard each other as peers with respect to the question of our share of the bill. Elga argues that, in such cases, suspension of judgment is called for, and it is in virtue of this feature that his view is standardly classified as a conciliatory view.

However, Elga emphasizes that his view does *not* entail that we should suspend judgment about all or even most controversial questions.[11] This is because

> in messy real-world cases . . . one's reasoning about the disputed issue is tangled up with one's reasoning about many other matters. As a result, in the real-world cases one tends not to count one's dissenting associates—however smart and well informed—as epistemic peers. (492)

That is, when one finds oneself in a disagreement about whether (e.g.) some proposed tax cut would stimulate the economy, that disagreement is typically not an *isolated* one. It is not as though one disagrees about that issue and nothing else; rather, one typically finds that one disagrees with the other person about any number of other, related issues as well (e.g. certain claims about human motivation and the ways in which people respond to financial incentives, or the effects of similar tax cuts that have been enacted in the past). And precisely because one takes one's interlocutor to be wrong about these surrounding, related issues, one will not regard him or her as equally likely to be right about the issue

[11] Indeed, in Elga's estimation, one of the most pressing challenges that any conciliatory view must overcome is what he calls *the problem of spineleness*: "the problem that an egalitarian view on how to respond to disagreement will recommend suspension of judgment on virtually all controversial issues" (492). Elga suggests that if a view about disagreement *did* recommend that one suspend judgment on virtually all controversial issues, then that would amount to a *reductio ad absurdum* of that view. So he is particularly adamant that the Equal Weight View has no such consequence. For criticism of Elga on this front, see McGrath (2007) and Kornblith (2010).

under dispute. That is, one will not count one's interlocutor as a peer with respect to the question of whether the proposed tax cut will stimulate the economy. Thus

> with respect to many controversial issues, the associates who one counts as peers tend to have views that are similar to one's own. That is why—contrary to initial impressions—the equal weight view does not require one to suspend judgment on everything controversial. (494)

But of course, *philosophy* is surely a paradigm of a domain in which the disputed issues are "messy" in Elga's sense: it is not as though typical philosophical disputes are isolated ones that take place against a background of generally agreed-upon common ground. (Consider just how messy things can get—and how quickly the messiness ensues—when disputes break out between Kantians and consequentialists, or between physicalists and dualists, or between epistemicists and supervaluationists.) In this respect, disagreements between philosophers are much more like disagreements about the probable effects of enacting some tax cut than the kind of disagreement that we find in the Mental Math Case. As far as Elga's conciliatory view is concerned then, such philosophical controversies are unlikely to require the dissenting parties to suspend judgment.

Consider next a move made by David Christensen in his (2011) defense of conciliationism. There, Christensen distinguishes between two different views that a conciliationist might endorse (15–16). On the first view, the *mere absence* of good reason to think that one is better informed or more likely to have reasoned correctly than the other person is enough to warrant belief revision in the event of a disagreement. On the second view, the mere absence of good reason to think oneself superior is insufficient to warrant belief revision; rather, belief revision is warranted when one has positive reasons to believe that the other person is (at least) equally well informed and equally likely to have reasoned correctly. Christensen argues that the conciliationist has compelling reasons to adopt the second type of view over the first. He also emphasizes that the second view is more limited in its scope, and hence, at least somewhat more limited in its potential to generate skeptical consequences. On the preferred picture

> when disagreement undermines one's rational confidence in some claim, the undermining must be based on one's beliefs about the other person. When those beliefs including extensive dispute-independent evidence of intellectual and evidential parity (as in the Mental Math Case), the undermining power of disagreement is high. But in some cases, one has little dispute-independent reason to be highly confident, one way or the other, about whether the other person is even one's approximate peer. In those cases, the undermining power of disagreement should intuitively be less. (16)

Now, although Christensen himself does not suggest anything like this, one might think that there is a crucial difference here between the Mental Math Case and typical disagreements between philosophers. In particular, the Mental Math Case is set up so that you and I have strong evidence that we are peers: specifically, there is the long history that supplies us with track-record evidence, track-record evidence to the effect that when disagreements of this kind have arisen in the past, each of us has turned out to be correct a more or less equal number of times. In contrast, there is typically *nothing like this* in cases of philosophical disagreement: if you and I find ourselves in a disagreement about some philosophical issue, we simply will not have the same kind of objective, non-question-begging

track-record evidence that would put us in a strong position to make a judgment about our comparative reliability.[12] (And this will be true even if we are very familiar with the other person's philosophical views, across a wide range of issues.) At best, it seems that what we really have to go on, at least in the first instance, are certain putative marks or indications of reliability—perhaps, for example, the ingenuity and resourcefulness that the other person manifests in parrying objections and devising new arguments for their view. But one might reasonably wonder just how strong the correlation is between the possession of such intellectual skills and reliable philosophical judgment, especially in view of the fact that the history of philosophy shows that such skills have been amply manifested by thinkers who agree with one another about very little (and so their possession is evidently compatible with a great deal of error). More generally, strong evidence that another person is your peer with respect to some philosophical question, or some cluster of philosophical questions, might be relatively hard to come by (see King 2012). And in the absence of such evidence, sticking to your guns might seem more defensible, as Christensen suggests about the general case. Of course, if hard evidence of comparative reliability is difficult to come by in the philosophical case, then it might be that you often end up lacking evidence that you are more likely to be correct than I am about some philosophical question. But as we have seen, on Christensen's version of conciliationism, merely lacking evidence that you are my superior is not yet enough to give you a reason to revise your opinion when we disagree; rather, what is relevant is whether you *have* good evidence that I am your approximate equal (or superior).

Thus, in somewhat different yet related ways, the work of Elga and Christensen suggests how one can endorse a view about disagreement that is recognizably conciliationary while denying that philosophical agnosticism follows as a normative upshot. Specifically, one might endorse the conciliationist idea that one is required to suspend judgment whenever one disagrees with someone whom one *takes to be a peer* (Elga), or with someone whom one *has good reason to think is a peer* (Christensen), while simultaneously emphasizing that such cases might turn out to be much less common than initial appearances suggest.

4. Egalitarian Versions of Conciliationism

Conciliatory views of peer disagreement call for a kind of epistemic democracy among peers, in some honorific sense of "peer" given by the theory. In determining what it is reasonable for one to believe about some question, one's own judgment is not privileged over the judgment of any other recognized peer. Given this, it is clear enough how a theorist who accepts a conciliatory view of peer disagreement might nevertheless end up denying that widespread philosophical disagreement warrants philosophical agnosticism: namely, by employing a notion of peerhood that makes it relatively difficult for philosophers who disagree with one another to count as peers, or to recognize each other as such. (Compare political systems that employ egalitarian, democratic decision-making procedures among

[12] On the significance of the (un)availability of non-question begging track record evidence in different domains, see especially McGrath (2009, 2011).

the class of citizens, but which are relatively stingy or sparing when it comes to granting citizenship status in the first place.)

Of course, this kind of maneuver immediately suggests possible conciliatory views that *do* seem to have skeptical upshot: namely, versions of the view that are relatively *inclusive* or liberal when it comes to counting disputing philosophers as "peers". Now, it is clearly possible to go too far in a liberalizing direction, even if one limits oneself to the class of active professional philosophers. For example, I have various half-baked opinions about how sentences containing problematic anaphora should be analyzed, but I would hardly expect philosophers of language working in the area to be moved by my opinions, and indeed, I would regard it as a kind of epistemological mistake on their part if they were. Such philosophers of language should not regard me as a peer, in any sense of "peer" that could possibly matter to the epistemology of disagreement.

In response, one might adopt some less inclusive construal of the notion, but one that is still inclusive enough to allow even philosophers who agree about relatively little of substance to count as "peers". For example, one might try something along the following lines: with respect to a given philosophical issue, count well-informed, apparently competent professional philosophers specializing in the area as peers, regardless of whether they agree or disagree with one another about the issue in question, and regardless of how much they agree or disagree about surrounding issues. Surely, one might think, if we adopted this or some similarly liberal construal of peerhood, and then combined it with a conciliatory view about the correct response to peer disagreement, this would require agnosticism when it comes to philosophically controversial issues.

But I think that even this thought cannot ultimately be sustained.

As a first step towards seeing why this is so, let's note the following point: in general, the fact that a given position can correctly be described as *controversial* among a population greatly underdetermines how widely accepted it is within that population. One way in which a view can be controversial within a population is for that population to be *evenly divided* about its truth, with an equal number of people on each side of the issue. But that is an unusual special case, and one that is seldom realized when dealing with populations of any significant size. In fact, a position can be controversial even if it is accepted by a significant majority of the relevant population. For example, over the past forty years, support for capital punishment among the American public has never dropped below 60%, and has often been significantly higher than that.[13] Nevertheless, during that same period, the legitimacy of capital punishment has often been cited—correctly, I believe—as a paradigmatic example of an issue that is controversial among the American public.[14]

[13] See <http://www.gallup.com/poll/1606/death-penalty.aspx> (accessed September 21, 2015).

[14] How popular can an opinion be and still be a controversial opinion? It is at least clear that when an opinion is held by *everyone* in a given population, it is uncontroversial within that population. But perfect unanimity is not required: the view that Adolf Hitler was a bad person is not controversial among the American public, even though there are some Americans who deny it. The general question of how substantial dissent must be in order for a majority opinion to count as controversial is likely both vague and context-sensitive. For our purposes, we need not pursue this question further; the important point is that a position can be controversial even if its supporters significantly outnumber its opponents, both in absolute terms and as a percentage of the relevant population.

Philosophy abounds with similar examples. Consider the following issues, each of which is central to a major subfield of the discipline:

(1) Epistemology: whether some of our knowledge is a priori
(2) Metaphysics: whether the best account of laws of nature is Humean or non-Humean
(3) Philosophy of science: whether scientific realism is true
(4) Philosophy of Mind: externalism *versus* internalism about mental content
(5) Philosophy of mathematics: Platonism *versus* nominalism about abstract objects
(6) Ethics: Cognitivism *versus* non-cognitivism about moral judgment
(7) Decision theory: whether one should take one box or two boxes in Newcomb's problem

Each of these issues is controversial within contemporary philosophy. Nevertheless, the best information that we have about the current state of informed philosophical opinion suggests that, in each case, such opinion comes down heavily on one side of the issue.[15] Thus, among epistemologists, 77% favor the view that there is a priori knowledge. Among metaphysicians, 72% favor non-Humean over Humean accounts of the laws of nature while fewer than 20% favor Humean accounts.[16] Among philosophers of science, 70% favor scientific realism while fewer than 10% favor anti-realism. Among philosophers of mind, 56% favor externalism about mental content while fewer than 20% are internalists. Similarly, 60% of philosophers of math favor Platonism about abstract objects while only 20% favor nominalism. Almost 75% of metaethicists favor cognitivism about moral judgment with approximately 12% favoring non-cognitivism. Among decision theorists, those who favor Two-Boxing outnumber those who favor One-Boxing by a forty-point margin (61.3%–21.3%). Many other examples could be provided, including the popularity of compatiblism about free will among metaphysicians, of the analytic–synthetic distinction among philosophers of language, of moral realism among metaethicists, of physicalism about the mind among philosophers of mind, of classical logic among logicians and philosophers of logic, of switching in the Trolley Problem among normative ethicists, and of commonsense realism about the external world among both all philosophers and epistemologists. While there are some controversial issues where informed philosophical opinion seems to be more or less evenly divided, such cases are much more the exception than the norm.

This empirical fact is worth bearing in mind whenever one looks to apply putative lessons from the epistemology of disagreement literature to the case of philosophy, for that literature typically focuses on cases of disagreement where opinion *is* evenly divided. Indeed, for heuristic reasons, the epistemology of disagreement literature often focuses on relatively idealized, two-person cases of disagreement. In these cases, one person arrives at

[15] See the results of the 2009 *PhilPapers* survey conducted by David Chalmers and David Bourquet at <http://philpapers.org/surveys/results.pl> (accessed September 21, 2015), which is the source for all of the numbers reported in this paragraph. The numbers reported here reflect the opinions of the survey's 'target faculty' (professors of philosophy at 99 leading departments) possessing the listed area of specialization.

[16] Here and throughout, the percentages fail to sum to 100 because the remainder of the respondents selected a third option, "Other", which covers such sub-options as "agnostic/undecided", "insufficiently familiar with the issue", and "the question is too unclear to be answered".

the view that p, while a peer arrives at the view that not-p, or at some view that is incompatible with p. Of course, any example that instantiates this basic structure is not only a case in which p is controversial in the relevant population, but is also a case in which opinion about whether p is perfectly divided within that population. Because the latter is a *very* special case of the former, we as theorists should be wary of generalizing plausible conclusions about such examples to cases of disagreement where opinion is not evenly divided.

Consider once again Christensen's Mental Math Case. Suppose that one shares the orthodox intuition about the example: that at the moment we learn that I arrived at the number 45 and you arrived at the number 43, we should suspend judgment about which answer is correct. Given this verdict about that particular case, consider two different lessons one might be tempted to draw by generalizing in different ways:

a) Once I learn that I have a peer who holds a view that is incompatible with p, then it is no longer possible for me to reasonably believe p: I should suspend judgment about whether p.
b) Once I learn that the distribution of opinion among the peers is evenly divided as to whether p, then it is no longer possible for me to reasonably believe p: I should suspend judgment about whether p.

Whatever arguments might be offered on behalf of b), it is clear enough that a) is false. According to a), learning that I have *at least one* peer who thinks that not-p is sufficient to undermine the rationality of my continuing to believe p. It thus entails that, in a case in which I simultaneously learn the opinion of multiple peers, the fact that one of the peers thinks not-p suffices to undermine the rationality of my believing p, even if all of the other peers independently arrived at the conclusion that p. But that is clearly incorrect, and is not something that any clear-headed conciliationist will accept. In the Mental Math Case, even if the rationality of your belief that each of us owes $43 is undermined when you learn that I arrived at $45 *and nothing else*, it is clearly not undermined by your learning that I arrived at $45 *and the other four people at the table also arrived at $43*. Indeed, in the latter circumstances the effect of learning the distribution of opinion within the group should have the effect of making you even more confident of your answer than you were originally, notwithstanding the fact that you now know that someone who is your peer arrived at an incompatible answer, something that you did not know before. The general moral, which no clear-headed conciliationist will deny, is that the numbers count: even if you know that your view is contradicted by some who are in as good or better position to pass judgment, this does not show that you should suspend judgment, for those who disagree might be outnumbered within the relevant group.

In fact, when one is dealing with groups whose members are at least slightly better than chance at answering a given kind of question, it is striking how quickly the probability that the majority opinion is correct increases with either the margin or the size of the group. Although the point is familiar from discussions of the Condorcet Jury Theorem, it's worth briefly illustrating with some numbers. Imagine a group of fifty people, each of whom is only slightly better than chance—say, 55% accurate—at answering a certain kind of 'yes' or 'no' question correctly.[17] In order to keep things parallel with the Mental Math Case and

[17] Of course, many canonical philosophical questions can be naturally cast in this form, e.g. "Is free will compatible with determinism?" or "Is any of our knowledge *a priori*?"

the other examples standardly used to motivate conciliatory views, we assume that each person arrives at his or her opinion independently. Of course, when the group splits perfectly down the middle, with twenty-five members of the opinion that p and twenty-five of the opinion that not-p, then the chance that either group is correct given just those facts is 50%. However, when the group splits 55/45, the chance that those in the majority are correct is approximately 73%. Suppose that we increase the size of the community from 50 to 70 while holding their reliability constant at 55%. In that case, when the group splits 55–45, the chance that the majority is correct is over 80%; if instead the group breaks 60–40, the chance that the majority is correct rises to over 94%.

The effect is even more dramatic when, instead of increasing the size of the original group, we increase their reliability even slightly. Suppose, for example, that rather than assuming that the members of the group are a slightly better than chance 55% reliable, we more optimistically assume that they are 60% reliable. On that assumption, when the group splits 55–45, the chance that the majority is correct is over 88%. When the group splits 60–40, the chance that the majority is correct rises to over 98%, with the chance that the minority view is correct less than 2%. In these circumstances, if we proportioned our subjective confidence to these chances, we would be virtually certain that the majority opinion is correct.

As noted above, there are many controversial philosophical issues with respect to which informed philosophical opinion divides quite unevenly, exceeding the 55–45 or 60–40 splits stipulated to hold in our toy models. Of course, even if we assume that an egalitarian theory of how to take the opinions of others into account is true, the high levels of confidence that would be licensed by lopsided majorities in the toy models would not be reliable guides to the levels of confidence licensed by similarly lopsided majorities within the philosophical community, for the toy models involve assumptions that clearly do not hold of the actual philosophical community. Notably, for example, each of the individuals in the toy model is stipulated to arrive at his or her views independently (as in the Mental Math Case), and it is manifestly not true that professional philosophers arrive at their philosophical views independently of one another. The toy models are thus meant to be suggestive rather than probative. Accordingly, the point that I wish to make is a relatively modest one: given the kinds of lopsided majorities that often obtain in philosophy, there is no reason to think that an adequate egalitarian theory (one which was sensitive to the various kinds of influence and dependencies in the way in which philosophers arrive at their views) would end up recommending agnosticism as the correct general response to philosophical controversy, as opposed to a relatively confident belief that a certain position is true in many cases.

Of course, we can imagine possible histories of how the actual distribution of philosophical opinion came about that *would* warrant agnosticism, given a conciliatory view of peer disagreement. For example, suppose that the following were true: not only do professional philosophers not arrive at their opinions in complete independence from one another, but in fact, any philosopher automatically inherits any view that is adamantly held by his or her dissertation adviser. The reason why over 75% of epistemologists think that there is at least some a priori knowledge is that they studied with the leading guru at The School of Rationalism, while the fewer than 15% who deny this studied with the leading guru at The School of Empiricism. Here the numbers should drop out entirely: assuming that the rationalist guru and the empiricist guru are counted as peers, the case effectively reduces

to the kind of two-person case of peer disagreement that is often the focus of attention in the literature, in response to which conciliatory views typically counsel agnosticism. But of course, we know that nothing like this fictitious history is actual.

We should also note that there are at least some *possible* conciliatory views that would warrant philosophical agnosticism as things actually stand. For example, consider a view according to which you cannot reasonably believe p so long as some formidable philosopher earnestly denies p. Given that with respect to almost any philosophically controversial issue, one can typically find at least one formidable philosopher on either side of the question, this conciliatory view would consistently recommend agnosticism in cases of philosophical controversy. But this conciliatory view is not a plausible one, as it in effect attributes too much undermining power to the views of formidable outliers.[18] (Even if some mathematically competent person in our group comes up with a different answer than the rest of us, it does not follow, on any remotely plausible view of disagreement, that suspension of judgment is required.)

Another worry about the line of argument sketched in this section runs as follows. In the toy models we have considered, it was assumed not only that the peers arrive at their views independently, but also that that they are at least slightly better than chance in answering the relevant questions correctly.[19] Is even this seemingly modest assumption too optimistic in the case of philosophy? Perhaps philosophers are so unreliable that they would be better off flipping coins than attending to arguments in trying to answer questions like "Is some of our knowledge a priori?" or "Is free will compatible with determinism?" In general, if it is assumed that the members of a group are sufficiently unreliable, then it is *not* the case that a lopsided majority is more likely to be correct than the badly outnumbered minority. Indeed, as is well known, in cases where the members of a group are anti-reliable, the Condorcet reasoning works in reverse, and the chance that the outnumbered *minority* is correct approaches one as the lopsidedness of the split increases. So if one has reason to believe that philosophers are sufficiently unreliable, then the mere fact that a substantial majority of the community has arrived at a particular view does not give one a reason to be confident that that view is true.

While this last point is correct, it is important to recognize that its truth poses no threat to the main thesis of this chapter. Again, that thesis is that there is no plausible view about the epistemology of disagreement that would require us to be philosophical agnostics. The possibility under consideration in this section is the following: if an egalitarian version of conciliationism turns out to be correct (that is, a version of conciliationism that is very liberal in counting people as "peers"), then we are required to be philosophical agnostics. But the crucial point is this: there is no plausible view about the epistemology of disagreement on which you are rationally required to conciliate your opinions with the members of a

[18] Compare: it is generally thought that the distribution of opinion among climatologists about whether man-made global warming is occurring increases, rather than decreases, the justification that most people have for believing that it is occurring, despite the fact that at least some of those best positioned to judge believe otherwise. (And despite the fact that the climatologists who believe that it is occurring did not arrive at that opinion in complete independence of one another.)

[19] According to an important extension of the Condorcet Jury Theorem due to Grofman et al. (1983), it is not necessary that every member of the group be at least slightly better than chance, only that the average of their individual probabilities be at least slightly better than chance.

group that is completely unreliable. For it can be shown mathematically that conciliating in such circumstances is a suboptimal way to manage your opinions: even if you are as unreliable as every other member of the group, you would be better off adopting a policy of simply sticking with your original opinion (Grofman et al. 1983).[20] The charitable way to understand conciliationist views, I believe, is to understand them as tacitly restricted to domains where the peers are (reasonably believed to be) at least marginally more reliable than chance with respect to the relevant subject matter. Call this the *Minimal Reliability Proviso*.

Suppose that some version of egalitarian conciliationism is in fact the correct account when the Minimal Reliability Proviso is satisfied, and that the Proviso is satisfied with respect to some philosophical questions. In these cases, the correct account of how we should take the opinions of others into account might very well recommend not agnosticism but rather a relatively confident opinion in the majority view. Of course, even if egalitarian conciliationism does not proscribe your holding a confident view about some philosophical issue, it might very well proscribe your holding your *own* preferred view about the issue (that is, the view that you would hold if you made up your mind on the basis of your own assessment of the issue).[21] One might try to downplay the significance of this by emphasizing the following point: in cases where a significant majority has emerged, it is only the members of the minority who would have to switch positions, so most of those with a view about the issue could go on believing what they already believed. However, even for members of the majority, the truth of egalitarian conciliationism would be of significance: for the fact that it is reasonable for them to continue holding the same view is *not* because this is the view that they judged most reasonable in the light of the considerations that bear on it, but rather because of its status as the view favored by the majority. For this reason, the truth of an egalitarian version of conciliationism would be a matter of great methodological significance for philosophy, even if such a view would not (as argued here) support widespread philosophical agnosticism.

[20] Informally, the point is well illustrated by an example suggested by Hud Hudson (personal correspondence). Imagine a confidence game taking place on a city street corner, in which a crowd of spectators are playing guess-which-cup-the-ball-is-under-this–time. There are five cups, one ball, a dexterous handler, and a lot of people losing money. The dexterous handler isn't cheating—he's simply very good at getting people to think that they know where the ball is when they don't. Suppose that five of us have been watching for some time now, and we each have the same dismal track record of 25%. After closely observing the latest shuffle, four of us report that it seems to us that the ball is under the far right cup, while it seems to you that the ball is under the far left cup. In these circumstances, you should *not* treat the fact that your peers have independently converged on a particular answer as evidence in favor of that answer and against your own; indeed, you should treat the fact that your peers have converged on a particular answer as evidence *against* that answer.

[21] As a referee pointed out to me: although in Elga's terminology "the spinelessness objection" refers to the concern that conciliationism will require one to suspend judgment on all controversial questions (cf. footnote 11), the term "spinelessness" perhaps more naturally suggests a different objection, namely, that conciliationism does not permit one to stand by one's *own* position, even in cases in which it does permit one to stand by *some* position.

5. Conclusion

While it is perhaps unsurprising that steadfast views make room for the possibility of rationally holding confident opinions about philosophically controversial issues, it is more surprising that plausible conciliatory views do so as well. Indeed, having argued for the chapter's main thesis, we can conclude by offering the following speculative conjecture, one that stands at variance with much thinking on this topic: *when it comes to philosophically controversial issues, those who faithfully adhere to plausible conciliationist views will often end up more confident than those who faithfully adhere to relatively pure forms of steadfast views*.

A natural assumption is that, when it comes to fiercely contested issues in philosophy, conciliatory views of disagreement will license at best relatively diffident, weak opinions (if not suspension of judgment), while steadfast views will be more hospitable to the holding of relatively strong, confident opinions. The underlying picture is that conciliatory views will lead to greater moderation in how one distributes one's credence among alternative positions in virtue of the weight that one is required to give to the views of others. In contrast, on a view such as the Right Reasons View, one can in effect ignore the opinions of others so long as one takes into account the arguments and considerations on which their opinions are based; freed from such potentially moderating influences and left to one's own devices, one ends up with very confident opinions. However, that picture is potentially misleading, both in what it suggests about conciliatory views as well as in what it suggests about steadfast views. First, a plausible conciliatory view will treat the distribution of peer opinion as mandating a middling credence *only if* opinion among the relevant group of people is more or less evenly split; on the other hand, to the extent that one position emerges as favored within the group, the conciliatory view will recommend increasingly greater levels of confidence in its correctness. Moreover, a view like the Right Reasons View will recommend a high degree of confidence in a given philosophical position only if the original, first-order evidence and arguments available to the believer really do decisively support that view, as opposed to a case in which the first-order evidence is mixed, with genuinely powerful arguments on both sides, in a way that recommends either suspension of judgment or some middling credence in the best-supported position.

Thus, the idea that conciliatory views will require relatively diffident opinions about controversial philosophical issues, while steadfast views will legitimate relatively strong opinions, depends on two background assumptions about what tends to be true in such cases: (i) that the first-order evidence, when taken by itself, is strong enough to warrant a confident opinion about the topic, and (ii) that creditable philosophical opinion is not only divided, but is divided more or less evenly. But on reflection, we seem to have relatively little reason to think these assumptions are generally true, or even that they hold in most cases.

With respect to (i): it is independently plausible that, in many cases of the relevant kind, the evidence that is available to the philosophical community (in the form of arguments, and so on) is relatively inconclusive, compared to the kind of evidence that one has for many of one's non-philosophical opinions, and is not of a kind that would license a particularly high level of confidence. After all, when it comes to controversial questions in philosophy, we frequently find formidable arguments and considerations pulling in opposite

directions, and even those arguments that seem compelling often lack the kind of probative force that is had by mathematical proofs, or even the kinds of considerations that can be brought forward in support of either our best-confirmed scientific theories or the more certain of our ordinary beliefs. In such cases, steadfast views that presuppose a permissive conception of rationality will see the "permissibility band" as encompassing a range of relatively middling credences, with levels of confidence approaching certainty squarely outside the band. Similarly, the Right Reasons View will not license a high level of confidence either, for the simple reason that such a level of confidence is not warranted by the original, first-order evidence.

With respect to (ii): based on the best information we have about the state of informed philosophical opinion, it seems that even when such opinion is divided, it is often *not* evenly divided: rather, in many cases, such opinion comes down more heavily on one side than the other. In cases in which informed philosophical opinion divides in a lopsided manner, plausible conciliatory views will often license the holding of relatively confident philosophical opinions (at least if there is a significant amount of independence in the way that those in the majority arrived at their view). This is because we can expect plausible conciliationist views to accommodate the following epistemic phenomenon: even when *no one* is in a particularly strong position to judge whether p, sufficient convergence on p can make it rational to invest a relatively high degree of confidence in p. For example, if each of us attempts to read a sign at a distance in poor viewing conditions, the fuzzy visual appearance that is presented to each of us might justify only a weak, tentative belief that the first word on the sign is "our" as opposed to "out". However, if it turns out that a substantial majority of our group weakly judges that the sign reads "our", then this might justify our investing a relatively high level of confidence that the sign says "our", despite the relatively impoverished character of what each of us had to go on in arriving at that view.

The sense in which conciliatory views counsel moderation should thus not be misunderstood. Conciliatory views will counsel one to moderate one's views in response to learning that others who are in a good position to judge think differently, but this is not to be confused with ending up with moderate opinions in cases in which competent opinion is unevenly divided. Indeed, it is possible that, when applied to the special case of philosophy, those who faithfully adhere to conciliatory views of disagreement will often end up more confident of their controversial philosophical opinions than those who faithfully adhere to more steadfast views.[22]

References

Armstrong, David (2006). "The Scope and Limits of Human Knowledge". *Australasian Journal of Philosophy* 84(2): 159–66.

Christensen, David (2007). "Epistemology of Disagreement: The Good News". *Philosophical Review* 116(2):187–217.

[22] Many of the ideas in this paper were presented at a meeting of my Spring 2013 graduate seminar at Princeton University, which was co-taught with John Hawthorne. I'd like to thank John and the students, as well as an anonymous referee for Oxford University Press, for very helpful feedback.

Christensen, David (2011). "Disagreement, Question-Begging, and Epistemic Self-Criticism". *Philosophers' Imprint* 11(6): 1–22.
Christensen, David and Lackey, Jennifer (2013). *The Epistemology of Disagreement: New Essays*. Oxford: Oxford University Press.
Elga, Adam (2007). "Reflection and Disagreement". *Noûs* 41(3): 478–502.
Feldman, Richard (2006). "Epistemological Puzzles about Disagreement". In Hetherington (ed.) *Epistemology Futures*. Oxford: Oxford University Press, 216–36.
Feldman, Richard (2009). "Evidentialism, Higher-Order Evidence, and Disagreement". *Episteme* 6(3): 294–312.
Feldman, Richard and Warfield, Ted (eds.) (2010). *Disagreement*. Oxford: Oxford University Press.
Frances, Bryan (2010). "The Reflective Epistemic Renegade". *Philosophy and Phenomenological Research* 81(2): 419–63.
Frances, Bryan (2013). "Philosophical Renegades". In Christensen and Lackey, 121–66.
Fumerton, Richard (2010). "You Can't Trust a Philosopher". In Feldman and Warfield, 91–110.
Goldberg, Sanford (2009). "Reliabilism in Philosophy". *Philosophical Studies* 124: 105–17.
Goldberg, Sanford (2013). "Disagreement, Defeat, and Assertion". In Christensen and Lackey, 167–89.
Grofman, Bernard, Owen, Guillermo, and Feld, Scott L. (1983). "Thirteen Theorems in Search of the Truth". *Theory and Decision* 15(3): 261–78.
Kelly, Thomas (2005). "The Epistemic Significance of Disagreement". In John Hawthorne and Tamar Gendler (eds.), *Oxford Studies in Epistemology, Volume 1*. Oxford: Oxford University Press, 167–96.
Kelly, Thomas (2010). "Peer Disagreement and Higher Order Evidence". In Feldman and Warfield, 111–74.
Kelly, Thomas (2013a). "Disagreement and the Burdens of Judgment". In Christensen and Lackey, 31–53.
Kelly, Thomas (2013b). "Evidence Can Be Permissive." In Steup et al., 298–313.
Kelly, Thomas (2014). "Believers as Thermometers". In Jonathan Matheson and Rico Vitz (ed.). *The Ethics of Belief* Oxford: Oxford University Press, 301–14.
King, Nathan (2012). "Disagreement: What's the Problem? or A Good Peer is Hard to Find". *Philosophy and Phenomenological Research* Vol. LXXXV(2): 249–72.
Kornblith, Hilary (2010). "Belief in the Face of Controversy". In Feldman and Warfield, 29–52.
Lasonen-Aarnio, Maria (2013). "Disagreement and Evidential Attenuation". *Noûs* 47(4): 767–94.
Matheson, Jonathan (2009). "Conciliatory Views of Disagreement and Higher-Order Evidence". *Episteme: A Journal of Social Epistemology* 6(3): 269–79.
McGrath, Sarah (2007). "Moral Disagreement and Moral Expertise". In Shafer-Landau (ed.), *Oxford Studies in Metaethics Vol. 4*. Oxford: Oxford University Press, 87–107.
McGrath, Sarah (2009). "The Puzzle of Pure Moral Deference". *Philosophical Perspectives* 23(1): 321–44.
McGrath, Sarah (2011). "Skepticism about Moral Expertise as a Puzzle for Moral Realism". *Journal of Philosophy* 108(3): 111–37.
McGrath, Sarah and Kelly, Thomas (forthcoming). "Are There Any Successful Philosophical Arguments?" To appear in an Oxford University Press *festschrift* for Peter van Inwagen edited by John Keller.
Rosen, Gideon (2001). "Nominalism, Naturalism, Epistemic Relativism". *Noûs* 35: 69–91.

Scanlon, Thomas (2014). *Being Realistic About Reasons*. Oxford: Oxford University Press.
Setiya, Kieran (2012). *Knowing Right from Wrong*. Oxford: Oxford University Press.
Steup, Mathias, Turri, John, and Sosa, Ernest (eds.) (2014) *Contemporary Debates in Epistemology*. 2nd edition. Oxford: Blackwell Publishers.
Titelbaum, Michael (2015). "Rationality's Fixed Point (Or: In Defense of Right Reason)". In John Hawthorne and Tamar Gendler (eds.), *Oxford Studies in Epistemology*, Volume 5. Oxford: Oxford University Press: 253–94.
van Inwagen, Peter (1996). "'It is Wrong Everywhere, Always, and for Anyone to Believe Anything on Insufficient Evidence'". In Jordan & Howard-Snyder (eds.), *Faith, Freedom and Rationality*. Savage, Maryland: Rowman and Littlefield, 137–54.
van Inwagen, Peter (2006). *The Problem of Evil*. New York: Oxford University Press.
Weatherson, Brian (2013). "Disagreements, Philosophical, and Otherwise". In Christensen and Lackey, 54–76.
White, Roger (2013). "Evidence Cannot Be Permisssive". In Steup et al., 312–23.

CHAPTER 21

FAITH AND REASON

LINDA ZAGZEBSKI

1. Introduction

When reason and faith are discussed together, it is usually because of the assumption, or at least the suspicion, that they can conflict in a deep and important way, and many writers interpret the alleged conflict in a way that is not flattering to faith. As Mark Twain allegedly quipped, "faith is believing what you know ain't so." I take for granted that if you know something ain't so, you have no business believing it. But it is not obvious that faith is believing what you know to be false, or reasonably believe to be false, or would believe to be false if you were being reasonable. In fact, it is not obvious how and whether faith conflicts with reason. In the next section I will begin by identifying the way in which faith and reason potentially conflict, concluding that faith has a component of belief on the word of God which does not conflict with reason directly, but which can be reasonable or unreasonable. In order to have a method for determining the reasonableness of a belief, I will distinguish two kinds of epistemic reasons, and will argue that trust in ourselves when we are epistemically conscientious is more basic than either kind of reasons, and it is more basic than any norms of reasoning. In the final section I will give my account of the place of faith in the epistemically conscientious person.

2. Can Reason and Faith Conflict?

Reason is usually thought to be a process or faculty that produces beliefs. Faith also is sometimes understood to be a process that produces beliefs. When we look at reason and faith this way, reason can, in principle, produce a belief that conflicts with a belief produced by faith. If that is all the conflict between faith and reason amounts to, there is nothing distinctive or even especially problematic about it. After all, we face conflicts between the products of different sources or processes for acquiring beliefs all the time. I may seem to remember something that conflicts with someone else's memory, or I may seem to perceive

something that conflicts with my memory, or I may draw an inductive inference that conflicts with my memory or a current observation, or someone else's memory. In all such cases, I have to decide which source of belief is more trustworthy on that occasion. I must do the same thing if reason generates a belief that conflicts with a belief generated by faith. The potential for such conflicts does not in itself raise any more difficulties than the potential for conflicts between perception and memory, or memory and induction.

I assume, however, that there are better and worse ways to handle these conflicts. Conflict need not be a special problem, but the problem would be serious if faith is an untrustworthy process, or if people tend to illegitimately favor faith whenever it appears to conflict with reason. That would be analogous to someone who illegitimately favors the product of their own memory over the testimony of other people, or a person who illegitimately favors current observations over past observations. We will need to address the way an epistemically conscientious person ought to think of faith, and how she handles conflicts of all these kinds, but the mere fact that conflicts can occur is not a special problem.

There is a modification of this way of framing the potential conflict between faith and reason that I think is more faithful to the tradition. Rather than to think of faith as something that produces beliefs, it is more common to think of faith as a state that includes belief as a component. On this view, faith is not a belief-producing process, but is the result of such a process, and that process is divine revelation. By revelation the believer comes to believe something on the word of God. As Francis Bacon remarked, "The knowledge of man is as the waters, some descending from above, and some springing from beneath; the one informed by the light of nature, the other inspired by divine revelation."[1] It is compatible with this approach that faith includes an emotional or conative component, but the potential for conflict with reason does not arise from anything non-cognitive, but from the possibility that the beliefs that are the product of revelation conflict with the beliefs that are the product of reason.

The view that faith is belief on revelation not only leaves aside the non-cognitive aspects of faith, but it also leaves aside the way in which faith can precede belief on revelation. Indeed, we sometimes speak of people who have faith in God but who do not base any beliefs on revelation. If such people have faith proper, its relation to reason differs from the relation between beliefs arising from revelation and beliefs arising from reason. It is faith in the latter sense that is the focus of this paper.

Does reason generate beliefs? That question can be interpreted in a number of different ways. If interpreted strictly, it is the question whether reason *all by itself* generates beliefs. The answer to that question is presumably yes, but the beliefs generated by reason alone are presumably very limited—beliefs in logical and mathematical truths, and analytic propositions. It is very unlikely that any beliefs in this category conflict with any beliefs putatively arising from revelation. If there is a conflict between faith and reason, it cannot be a conflict between beliefs based on revelation and beliefs arising directly from reason in the strict sense.

But suppose we broaden the notion of a belief based solely on reason to include beliefs in propositions allegedly known a priori but which have substantive content. Examples would be the belief that a cause must precede its effect, or the belief that a physical object cannot

[1] Bacon 2000 [1605], Bk. VII, ch. V, sec. 1.

exist in two widely separated places at the same time, or the belief that an object cannot go out of existence and then come back into existence at a later time. Perhaps these beliefs are in the category of the synthetic a priori. Of course, the status of these beliefs is disputable, but a case can be made that they arise from reason alone, yet have content that is about the world. These examples are more promising as cases of potential conflict between a belief based on revelation and one based solely on reason.

Consider, for example, the following set of propositions that, allowing for some vagueness, are inconsistent:

B1. Human beings will rise from the dead some time after death.
B2. A human being ceases to exist at death (or very shortly thereafter).
B3. An object cannot go out of existence and come back into existence at some later time.

Someone might believe B1 as a belief of faith. That is, she might take B1 to be the product of divine revelation. She might also think that B3 is a deliverance of reason in the expanded sense I just described. Of course, there is no conflict without B2, but that proposition might be put in the category of common sense. If so, the conflict among B1, B2, and B3 is a conflict among a belief of faith, a deliverance of reason in the expanded sense, and a belief of common sense. This case is the closest I can think of to one in which a belief that is putatively a direct deliverance of reason (B3) conflicts with a belief of faith (B1), but even this case is not a direct conflict since the conflict requires B2. Notice also that although the conflict needs to be resolved, it is not obvious how it should be handled. There are many things we could say about the basis for all three of these beliefs. For instance, although B2 and B3 are reasonable beliefs, must we say that reason *demands* B3, or that common sense *demands* B2? Likewise, B1 may be a belief of faith, the result of a process of communication from God, but is it obvious that B1 is the correct belief to form as the result of the revelatory process?

So far I do not think we have an example of a direct conflict between a requirement of reason and a requirement of faith, but there is an even broader notion of reason that would explain the perceived conflict. When people think of reason and faith as conflicting, they often have in mind a notion of reason according to which we say that some beliefs are reason*able* and others are not. In the broadest sense of reason, some beliefs are not reasonable even though they are not deliverances of reason alone, and certain other beliefs are unreasonable even though they do not contradict the deliverances of reason alone. In this sense we say it is reasonable for me to believe that I have made mistakes and will make mistakes in the future. It is reasonable for me to believe that I should save money. It is not reasonable for me to believe that I will win the lottery. In general, it is not reasonable for me or anyone else to form beliefs out of wishful thinking, or out of stubborn refusal to change, or out of hatred of the source. In this sense of the reasonable, B2 and B3 are reasonable beliefs, and they conflict with B1. In this sense of the reasonable, of course, a reasonable belief can conflict with another reasonable belief or set of beliefs. It is reasonable for me to believe that I parked my car in the first row of the parking lot, based on my memory of doing so, but it is also reasonable for me to believe that I parked in the second row, based on the testimony of my trustworthy research assistant. As I have said, conflicts need to be resolved, but there is nothing especially problematic about such cases. It could turn out that B1, B2, and B3 are all reasonable even though they conflict and the conflict must be resolved.

We are getting closer to the kind of conflict that can arise between faith, or belief based on revelation, and reason in the broadest sense, which I have interpreted as the requirement of reason to believe reasonably. Is it reasonable to believe what people tell me? Belief on revelation is a species of believing what people tell me. We think that it is often reasonable to believe what we are told, but sometimes it is not. Maybe the person I take to be the source is not trustworthy and if I was reasonable, I would be aware of that fact, or maybe the putative source does not even exist. I am not reasonable in believing my research assistant's testimony about the location of my car if I am not reasonable in trusting my research assistant about this matter, or if I do not reasonably believe that my trustworthy research assistant is the testifier. If I do not have a research assistant and am unreasonable in believing that I do, then I do not reasonably believe that my research assistant is the testifier. Similarly, faith can conflict with reason if I am unreasonable in believing what I take to be a revelation from God. I am unreasonable if I am not reasonable in trusting God, or if I am not reasonable in believing that the source of the testimony is God. If I unreasonably believe that God exists, I am not reasonable in believing that God is the source of the testimony. In discussing the reasonableness or unreasonableness of faith, we will need to address the conditions that make believing what I am told reasonable, and the conditions that make it unreasonable, and we will need to say something more about the sense of reasonable that is intended.

The most general thing we can say about the reasonable is that it is doing the right thing by way of believing, and being unreasonable is making a mistake in one's believing. One way I can make a mistake is by violating a rule of reasonable belief. Similarly, one way of doing the right thing epistemically is by following a rule of reasonable belief. But there are other ways of making a mistake than by violating a rule, and there are other ways of forming a belief in the right way than by following a rule. If I form a belief out of wishful thinking, I am not violating a rule of reason even though I am doing the wrong thing epistemically. If I form a belief out of proper trust in others, I am doing the right thing epistemically even though I am not following a rule. If someone claims that faith conflicts with reason in that faith includes belief that is unreasonable in this broad sense, the claim must be that forming a belief based on revelation is an epistemic mistake. It is doing the wrong thing by way of believing.

Here we must make a qualification. What it is reasonable *for me* to believe might not be the same thing as what it is reasonable for someone else to believe. My own experiences, memories, intuitions, and the results of my past reflections on those conscious states are mine alone. They give me, but not someone else, a kind of reason for belief that others lack. I will turn to this issue in the next section, and will argue that there are two kinds of epistemic reasons, one of which is irreducibly first personal. I will then argue that the primary norm of belief formation is my own epistemic conscientiousness, trust in which is a first person reason for belief. I have the property of conscientiousness when I do my best with the faculties I have to reach epistemic ends. The broadest sense of reason is being reasonable. Being reasonable is the same thing as being epistemically conscientious, and any violation of epistemic conscientiousness is unreasonable.

3. Two Kinds of Epistemic Reasons

Let us look at how we might devise a way to tell if a belief is reasonable—if a person is reasonable in having it. I have assumed that we can identify some reasonable beliefs and some unreasonable beliefs, and I gave some examples of each. I assume we can also identify some of the things reasonable persons do. One of the things reasonable persons do is to have epistemic reasons for their beliefs. What I mean by an epistemic reason is something on the basis of which it is reasonable for someone to settle for herself whether *p* in so far as her goal is truth. I assume that reasonable persons desire truth, and having a reason for a belief is a means to satisfying that desire. An epistemic reason need not be sufficient to settle the question whether *p*, but it is the sort of thing that can do so, normally in conjunction with other epistemic reasons. My position is that there are two kinds of epistemic reasons—one third personal, the other irreducibly first personal.[2]

What I call *theoretical reasons* for believing *p* are facts that are logically or probabilistically connected to the truth of *p*. They are facts (true propositions) about states of the world that, taken together, give a case for the fact that *p*. Theoretical reasons are not intrinsically connected to believing, but they are reasons because a reasonable person who comes to believe them and grasps their logical and probabilistic relations to *p* will see them as indicating the truth of *p*. Theoretical reasons can be shared with others—laid out on the table, so they are third personal. They are relevant from anyone's point of view. The connections between theoretical reasons and what they are reasons *for* are among the facts of the universe. What we call evidence is most naturally put in this category of reasons, but the notion of evidence is multiply ambiguous, and the nature of evidence is not crucial to the distinction I am making.

In contrast, what I mean by *deliberative reasons* have an essential connection to *me and only to me* in my deliberations about whether it is the case that *p*. Deliberative reasons connect me to getting the truth of *p*, whereas theoretical reasons connect facts about the world with the truth of *p*. Like theoretical reasons, deliberative reasons are *reasons* because a reasonable person, one who cares about the truth, takes them to be indicators of truth. Deliberative reasons provide me reasons for *p* that are not simply weightier than the reasons they provide others. They are not reasons at all for other persons. They are irreducibly first personal.

To see the distinction I have in mind, consider experience as a reason for belief. If I have an experience, the *fact* that I have it is a theoretical reason that supports a variety of propositions. I can tell you about my experience, and if you believe what I tell you, you can then refer to the fact that I had the experience as a reason to believe whatever it supports. So suppose I have the experience of seeming to see a scissor-tailed flycatcher near my home. You and I can both refer to the fact that I had that experience as a reason to believe that the

[2] Parts of this section on the distinction between the two kinds of reasons appear in similar forms in Zagzebski 2012a and in Zagzebski 2012b. I argue that the distinction can be used to illuminate a variety of problems in philosophy of religion in Zagzebski 2011. In Zagzebski 2015, I apply the distinction to the nature of faith, and in Zagzebski 2014, I apply the distinction to the infinite regress of reasons.

flycatchers have not yet migrated, and so can anybody else who is aware of the fact that I had the experience. The fact that the experience occurred is therefore a theoretical reason. It is on the table for anyone to consider, and anyone can consider its logical and probabilistic connections to other facts about the world.

However, I am in a different position with respect to my experience than you are, because I not only grasp the fact that I had the experience; in addition, I and I alone *had* the experience. *I* am the one who saw the flycatcher with its long split tail. That visual experience can affect my reasoning processes, emotional responses, and the way I come to have or give up certain beliefs directly, and that is perfectly normal for human beings. In contrast, the fact that I had the experience is something you and I and any number of other persons can come to believe. So my experience of seeing the scissor-tailed flycatcher gives me a reason to believe that the birds have not yet gone south. You cannot have my experience, but you can believe that I had the experience. When you do so, you are not accessing my experience; you are accessing the fact that the experience occurred. Of course, I can access the same fact, but my having a reason to believe that scissor-tailed flycatchers have not yet migrated does not depend upon my accessing the fact that I had the experience of seeing one. The seeing itself gives me a reason to believe they are in the neighborhood.

Another type of deliberative reason is what are often called intuitions in one of its senses. What I mean by an intuition is, roughly, something internal to the mind that responds with an answer to a question, often about a concrete case. For example, I have the intuition that it is not morally permissible to directly kill an innocent person to save five others, but someone else might have a different intuition. Most philosophers have the intuition that a Gettier case is not an instance of knowledge, but we probably have encountered students who do not have that intuition. I have no position on the strength of an intuition as a reason to believe what the intuition supports. Maybe it is strong, maybe it is not. But in so far as it is a reason at all, it is a deliberative reason. My intuitions are mine alone, and they give me, but not you, a particular kind of reason for certain beliefs. But again, the fact that I have an intuition can be put out on the table. I can tell you that my intuition is such and such. When I do so, I give you a theoretical reason supporting some position. The fact that many people have the same intuition can also be used to support a position. So the fact that a large majority of those persons who have carefully thought about the nature of knowledge have the intuition that a Gettier case is not an instance of knowledge supports the position that a Gettier case is not an instance of knowledge. Intuitions, then, are like experiences. An intuition and an experience provide the agent with first person reasons to believe something, but the fact that the experience occurred or that the intuition is what it is can be treated as evidence, as a theoretical reason for the truth of some proposition.

Experience and intuition reveal an important feature of deliberative epistemic reasons: They are psychic states of a person that seem to him or her to indicate the truth of some proposition *p*. Human beings are constituted in such a way that certain states are like that.

We would expect, then, that other psychic states can have the same function, for instance memories and certain emotions. But my purpose is not to give an exhaustive list of the kinds of conscious states that can be deliberative reasons for beliefs. What is relevant for my purposes in this chapter is that reasonable persons take certain of their conscious states to be indicators of the truth of some proposition. I propose that these states include

experiences, memories, intuitions, and some emotions. What makes these states reasons is that a reasonable person takes them to be truth-indicators. What makes these reasons deliberative is that they are reasons only for their possessors.[3]

4. The Primacy of Epistemic Conscientiousness

What happens when reasonable persons reflect upon their reasons? Let us start with theoretical reasons. If I reflect upon those reasons, I realize that my taking a certain set of facts as reasons for *p* is not sufficient to make it likely that *p* is true. That is because my taking something to be a set of reasons for *p* is irrelevant to the actual connection between those reasons and *p* unless I am taking them properly—have accurately identified the appropriate facts, have figured out the correct logical and probabilistic relations between those facts and *p*, have appreciated the significance of individual facts, and have not left anything out. Hopefully all of that is a fact about me. But it would not be a fact about me unless it is a fact that I have the kind of powers that enable me to find out facts about the world. But now the question arises: How can I access *that* fact?

It has been pointed out by others—for example, Keith Lehrer (1997), Richard Foley (2001), and William Alston (1986, 2005) that any reasons I have to believe that my powers connect me to the truth are circular. I cannot tell that my powers get me to the truth without using those powers. In fact, I cannot tell in a non-circular way that my epistemic powers ever get me to the truth, much less that they get me to the truth reliably. It would take a perspective outside of my mind to identify the quality of the relationship between my mind and a world outside of it, but I am a being who can never ascend to a perspective outside my own mind, save in imagination.

I have argued that deliberative reasons are not facts, but are conscious states that reasonable persons take to indicate the truth of some proposition. Notice that I face the same problem for the way conscious states get me to the truth as I do for the way theoretical facts get me to the truth. Just as I have no way of telling in a non-circular way that my attempt to access the facts gets me to the truth, I also have no way of telling that my sense experience, memory, or intuition gets me to the truth without using some of my own powers. I cannot tell that any of my conscious states connect me to the truth without using powers the truth-conduciveness of which I cannot determine without appealing to particular outputs of those same powers.

What is the reasonable response to the problem of the circularity of reasons? Both Foley and Alston maintain that it is epistemic self-trust, and I agree with them about that. It is reasonable to think that reasonable persons generally do not respond to reflection upon the circularity of reasons by becoming skeptics. They believe or take for granted that their

[3] As I have argued in the places I have cited it is important to distinguish theoretical from deliberative reasons for at least two reasons. One is that the two kinds of reasons do not aggregate. The other is that I have no control at all over the theoretical reasons, but I have the control of an agent over deliberative reasons.

faculties are generally conducive to getting them the truth, and I think that is part of epistemic self-trust. So what I mean by epistemic self-trust includes a belief component. I think that it also includes an affective component that is opposed to doubt. My position is that trust is in part a feeling as well as a belief. It dispels doubts or holds them at bay. It is reasonable to believe that reasonable persons have self-trust in this way, and therefore, it is reasonable to have self-trust in this way.

Notice also that if a reason for *p* is something on the basis of which a reasonable person settles for herself whether *p*, then self-trust *is* a reason, but it is a deliberative reason, in the category of experiences, memory, and intuitions as a ground for belief. It also follows from what I have said that epistemic self-trust is more basic as a reason to believe any proposition *p* than any other reason I can have, whether theoretical or deliberative.[4] The most basic epistemic reason I have is a reason for me and me alone.

When a person reflects, she thinks that her trustworthiness is greater if she summons her powers in a fully conscious and careful way, and exercises them to the best of her ability. What I am calling *epistemic conscientiousness* is the state or disposition to do that.[5] Conscientiousness is important because we do not think that we are equally trustworthy at all times. We trust that there is a connection between trying and succeeding, and the reflective person thinks that there is a closer connection between trying with the full reflective use of one's powers, and succeeding. Conscientiousness comes in degrees. There is a probably a degree of conscientiousness operating most of the time since we have some minimal awareness of ourselves and the exercise of our powers most of the time. But higher degrees of conscientiousness require considerable self-awareness and self-monitoring.

It follows from what I have argued that there are two levels of self-trust, both of which are more basic than any reasons or evidence we can identify. First, there is the general trust in our faculties that I argued is a reasonable response to epistemic circularity. Second, there is the particular trust we have in our faculties when we are epistemically conscientious––exercising our truth-seeking faculties in the best way we can.

Trust in ourselves when we are conscientious leads to another deliberative reason for belief: trust in others. Trust in others is natural, but reflection shows that it is rationally required. My position is that if, in believing in a way I trust, I come to believe that others have the same powers I trust in myself, then given the a priori principle that I ought to treat like cases alike, I am rationally required to have a basic trust in their powers in the same way I have a basic trust in my own. If I reasonably trust their powers, I have a prima facie reason to believe the product of the exercise of their powers.[6] Since I trust myself in particular when I am conscientious, it follows that when I conscientiously believe that another person is conscientious in some domain or aspect of belief-formation, I owe him the same particular trust that I have in myself when I am conscientious.[7]

[4] I do not mean to suggest that self-trust is a more basic fact of the universe than theoretical reasons. But self-trust is more basic for me in my deliberations about what to believe than my use of theoretical reasons.

[5] Note that as I define conscientiousness, it does not have any relation to duty.

[6] Some philosophers prefer the term "*pro tanto* reason" for what I call a "prima facie reason." By a prima facie reason I do not mean something that is simply an apparent reason, or something that can be a reason for a while and then lost. If something is a prima facie reason for p, it counts in favor of p even if it is ultimately defeated or outweighed by other reasons.

[7] I defend the argument of this paragraph in some detail in Zagzebski 2012a, ch. 3.

My acceptance of epistemic rules of all kinds depends upon trust in myself and others. It is because of trust in self and others that I trust the rules of reasoning that have been devised by persons I trust. The intellectual virtues are the qualities of conscientiously reflective persons who attempt to get truth or some other epistemic good such as understanding. It is reasonable for me to attempt to acquire the intellectual virtues only because it is reasonable to have a more basic trust in the epistemic powers of cognitive agents. There would be no point in rules for the conscientious exercise of powers that are devised by the conscientious exercise of those same powers unless we reasonably trusted the powers. Similarly, there would be no point in advocating open-mindedness, intellectual carefulness, courage, etc, each of which is a trait that we identify as desirable for epistemically conscientious agents by the conscientious exercise of the powers of conscientious agents, unless it was reasonable to trust those powers. The trust that undergirds epistemic rules and the identification of the intellectual virtues is first personal. There is therefore a deliberative epistemic reason that is not only more basic than theoretical reasons and other deliberative reasons, but it is more basic than the norms of belief-formation.

Trust in others leads to trust in communities, some of which consist of living persons (e.g., members of one's academic profession), and some of which extend far into the past. Religious communities are almost always in the latter category. There are deliberative reasons for the beliefs of a community as well as theoretical reasons, and the way the community identifies theoretical evidence often depends upon trust in the community's ability to get the truth, just as it does for individuals. Of course, members of a community believe many things that do not arise out of the community per se, but communities often function like a self: they have norms of belief formation and shared background beliefs that derive from trust in the community that operates the way self-trust operates for an individual.

I have said that epistemic self-trust is reasonable because reasonable persons have it. Reasonable persons also have basic trust in others. It might appear that I am suggesting that reasonable persons trust out of necessity and out of blind hope, which reasonable persons themselves ought to take to be unreasonable. But that is not correct. It is not blind hope to manage the self as all reasonable persons do. The self manages itself by reflecting upon its conscious states. That is what a self does in a world that is much larger than the self and which includes many other selves. Perhaps we wish that that were not the case. We might prefer to be able to go outside our self in order to determine which states of the self ought to survive and which ought to be given up. But we know that is impossible. A self just is a being that can only manage itself from the inside.

My view of the primacy of conscientious self-reflection has the consequence that ultimately our only test that a belief is true is that it survives conscientious reflection, including future reflection on experiences we have not yet had, and future judgments about the past and present. The way I identify what points to the truth is the same as the way I identify what the truth is. All I can do is to use my faculties the best way I can, that is, to be conscientious, and since future conscientious reflection includes reflection on the products of past conscientious reflection, survival of reflection in the future is my only way to tell that my past reflections have gotten me to the truth. Of course, I am not proposing that truth *is* what survives conscientious self-reflection. If it were, we would not need trust in conscientious self-reflection. But conscientious self-reflection is basic.

In summary, we think that reason is a faculty that delivers beliefs that aim at the truth, and our ways of forming beliefs can be reasonable or unreasonable. We think that reason determines norms for belief-formation, and some of those norms are not rules, but intellectual virtues that reasonable persons desire if they want the truth. We also speak of reasons for belief– something on the basis of which a reasonable person takes some proposition to be true. I distinguished two kinds of epistemic reasons, one third personal, the other irreducibly first personal. I argued that trust in our powers when we are conscientious—using them as best we can in order to get truth, is a first personal reason that is more basic than any theoretical reasons and any other deliberative reasons. It is more basic than the norms we identify as norms of reason. It is more basic than anything we call reasons or reasonable ways of believing. Being reasonable is just being epistemically conscientious.

5. Testimony, Faith, and Conscientious Belief

So far I have argued that the potential conflict between faith and reason reduces to the issue of whether the belief component of faith is reasonable. I then argued that being reasonable in belief is being epistemically conscientious. It is on the basis of epistemic conscientiousness that we adopt norms of reason and take certain propositions or conscious states to indicate the truth of a belief. It follows that faith is reasonable just in case the belief component of faith is epistemically conscientious. Ultimately, our only way to tell that the belief is true is that it survives our own conscientious reflection, or reflection upon our total set of beliefs with the aim of getting the truth.

Let us look next at how the belief component of faith arises. Faith is belief based on revelation, which means it is belief based on a type of testimony. To determine the reasonableness of faith, then, we will need some account of the nature of testimony. That is, of course, a hotly disputed topic, but I think that the distinction between the two kinds of epistemic reasons helps us understand the dispute, and I believe the distinction reveals an important feature of faith.

Some philosophers maintain that it is reasonable to believe testimony only when one has evidence that the testifier is reliable. That is to say, we are reasonable in believing testimony only if we have theoretical reasons to believe the testifier is reliable. Other philosophers maintain that treating testimony as evidence ignores what is most crucial to the testimonial act. According to the latter view, the speaker presents her testimony as providing a reason to the hearer to believe *p*, and she intends that the hearer believe what she says.[8] The speaker is asking for trust. The hearer may give the speaker trust and thereby come to believe what the speaker tells her.

The trust model of testimony seems to me to be basically correct. But it raises the following puzzle. How can the process of asking for trust and giving it provide a *reason* for me to

[8] This view of meaning appears in Paul Grice's classic paper, "Meaning," *Philosophical Review*, vol. 66, 1957, pp. 377–88, reprinted in Grice, *Studies in the Way of Words*, Harvard University Press, 1989. Richard Moran (2005) refers to it in defending a view of testimony similar to the one I defend here, calling it the "assurance model." Benjamin McMyler (2011) nicely defends a similar view of testimony, calling it the "second person" model.

believe something? It does not seem to be in the right category to be a reason. One would think that all that matters is that I have theoretical reasons to believe that the speaker believes *p* and is reliable. In that case I have good evidence that *p* is true. How can it add to my evidence to learn that in addition to her believing *p*, the speaker also has the intention that I should believe *p* and asks me to trust her? My recognition of the speaker's intention appears pointless.[9] In fact, any feature of the speaker's attitude towards me or my attitude towards her seems to be pointless.

My answer to the puzzle is that trust requested and given does not add to my theoretical reasons, but it gives me a deliberative reason, the kind of reason that only makes sense from the point of view of a particular person deliberating about what to believe. My relation to the testifier is essential to the grounds for believing what she tells me. Features of the testifier that only persons can have provide my reasons for belief: taking responsibility for believing conscientiously, and intending that I trust her in carrying out that responsibility. The testifier exercises the control of an agent over those reasons. I exercise the control of an agent in accepting them as reasons for my belief, which means that they are in some sense voluntary. These reasons are entirely different from those provided by a reliable instrument such as a thermometer. I do not deny that the testifier *can* be treated like a thermometer. I might simply judge that she is reliable in the relevant domain and has a belief in that domain, and therefore, I conclude, it is likely to be true. In this way I have theoretical reasons to believe what she tells me. Those reasons are reasons for anybody to believe what she says. They are third person reasons. In contrast, my trust in the testifier's responsibility in getting to the truth of *p* conscientiously for both of us is a reason that is essentially connected to the relationship between the testifier and myself. I feel let down if what the speaker says turns out to be false, and the speaker feels aggrieved if I do not believe her. I would have no reason to feel let down if what the speaker tells me is merely evidence, and she would have no reason to feel aggrieved when I do not take her word that *p* if all she is doing is providing me evidence for *p*.

On what grounds should I trust the speaker? I might have higher order theoretical reasons to think she has theoretical reasons to believe *p*, but that is not sufficient to trust her, although it might be sufficient to believe what she says. That was the point of my claim that it is possible to treat another person as a source of information on a par with a device like a thermometer or a calculator. But to trust a person, I need deliberative reasons, reasons that give *me* in particular a reason to trust *her* in particular. These reasons require reliance upon her attitude towards me as well as her epistemic abilities in the domain in question. These reasons are deliberative since they involve a relationship between the testifier's intention and my acceptance of her intention with regard to me.

The evidence view of testimony cannot explain these features, but we can see why the evidence view of testimony exists. It is the view we are forced to have if we think the only kind of reasons are theoretical reasons. If the only kind of reasons are theoretical, revelatory belief must either be based on theoretical reasons or properly basic. Plantinga (1983) famously argued for the second option. Locke and others have argued for the first.[10] I think

[9] This puzzle for the evidence model of testimony is raised by Moran (2005), p. 15.

[10] Locke argues for this position in more than one place. See his *Essay*, Bk. IV, ch. 18, sec. 7.

that revelatory beliefs are best understood as cases of believing God as described by the trust model of testimony. We have deliberative reasons for believing what God tells us.

If we follow this model, when God tells me that *p*, God takes responsibility for the truth of *p* for me and for all other intended recipients of his revelation. God intends that I believe him, and he acknowledges that we who are the recipients place epistemic trust in him by believing him. On the model I am proposing, then, religious faith is believing God, as Elizabeth Anscombe (1979) argued long before the current dispute about the nature of testimony came into the literature. The ground of faith is trust in God, which gives me a deliberative reason to believe what God tells me. As in ordinary cases of testimony, the testifier is responsible for the justification of the content of the belief, and the recipient is responsible for having appropriate trust in accepting the testimony. Since we have the control of an agent over deliberative reasons, and trust is a deliberative reason, an important consequence of this model of faith is that it explains how the belief component of faith can be in some sense voluntary.

Believing a person who is currently speaking to me or who has written a book or sent me an email is not very mysterious, but believing God requires a theory of revelation to explain how communication between God and me can succeed. There are a number of accounts of revelation compatible with the view of faith I am endorsing here. Some focus primarily on the historical witness of an originating event, such as the revelation to Moses on Sinai, which is passed on to us in an unbroken chain. Others emphasize the importance of oral tradition in transmitting revelation and its mediation through one's religious community. Others emphasize the direct action of the Holy Spirit on the believer on the occasion of coming to faith, as Plantinga (2000) does. All of these models are compatible with the view that faith is a relation between two persons, God and the believer. God speaks to us, the communion of the faithful, or directly to me, but either way, what I believe depends upon my relation to God as the ultimate testifier of the content of my belief, and it is upon him that I place my trust for the truth of what I believe.

There are Christian versions of all three models, and in each one trust in the Holy Spirit is needed. It is obviously needed in the third model, and it is needed in the first two since it is necessary to trust the accuracy of the transmission of revelation and the way that revelation is mediated through the Church. Throughout the Christian era there have been authorities with the responsibility to preserve and pass on the deposit of faith, so trust in them is necessary, but trust in a long succession of human beings is not likely to be strong enough to give a person confidence that she has been faithfully given Revelation from God without trust in the guidance of the Holy Spirit.

Notice also that the view that revelation is the direct action of the Holy Spirit on the believer can be used as part of a model that belief produced by revelation in that manner is properly basic, or it can be used as part of a model of revelation as testimony, justified by the recipient's trust in the testifier. If the Holy Spirit can produce beliefs, he can also produce trust, but the difference is that the model of faith as trust in testimony includes responsibility on the part of the recipient to trust reasonably. If I conscientiously judge that trusting the Church is trusting God, and that taking beliefs from God mediated through the Church will produce in me beliefs that survive present and future conscientious self-reflection, my trust is reasonable and responsible. On the view of faith I am

proposing, then, even though faith is a gift of the Holy Spirit, it is reasonable in the same way trust of any kind is reasonable. Indeed, it is reasonable in the same way beliefs of any kind are reasonable—they are formed and maintained in a way that is epistemically conscientious. There is no more that human beings can do but to use our faculties in the best way we can to reach their end, which in the case of our epistemic faculties is the truth.

When I discussed the potential conflict between reason and faith at the beginning of this chapter, I said that there is more than one way it can be unreasonable to form a belief on testimony. The belief would be unreasonable if it is unreasonable for me to trust the person I take to be the testifier, or if it is unreasonable for me to believe that the person I take to be the testifier is the testifier. If it is unreasonable for me to believe the person I take to be the testifier exists, then it is unreasonable for me to believe that the person I take to be the source of the testimony is the testifier. So the following beliefs must be reasonable if my belief that *p* based on revelation is reasonable: (1) God exists, (2) God is the testifier of *p*, and (3) God is trustworthy with respect to testimony that *p*. I think that it is reasonable to think that (3) is reasonable if (1) and (2) are reasonable, so the focus of attention in the issue of the reasonableness of faith is the reasonableness of (1) and (2). Given what I have said, these beliefs are reasonable just in case they are formed and maintained in an epistemically conscientious way.

Of course, (2) cannot be reasonable unless (1) is, and a conscientious believer whose total set of reasons, both theoretical and deliberative, makes it unlikely that (1) is true, will believe it is even less likely that (2) is true, but that does not necessarily mean that anyone's belief (2) depends upon a prior or more fundamental belief (1). As Sandra Menssen and Thomas Sullivan have argued (2002), the case for the truth of (1) can be enhanced by the case for the truth of (2). That is because evidence of communications of a certain kind can make it probable that a being of a certain kind exists. For instance, the *Iliad* is evidence of the existence of an author with certain properties, whom we get to know through his works. For some conscientious believers, the relationship between belief (1) and belief (2) is like that. Given their total theoretical evidence and relevant deliberative reasons, belief (1) can gain credibility from the reasons for (2). Clearly, for others, there are more reasons for (1) than for (2). And for many conscientious believers, beliefs (1), (2), and (3) are components of a set of beliefs of a religious community, integrated with moral values, emotions, and rituals that interpret their lives and give them meaning. Trust in particular beliefs is often derivative from trust in the community, and trust in the community is partly, but not wholly, epistemic trust. That is, trust in the community may be trust for the sake of goods in addition to truth.

Wisdom is a quality that combines knowledge of important truths with moral qualities and the ability to guide and counsel others. Trust in a religious community is often trust in the wisdom that resides in the community. Epistemic trust in it is an aspect of trust in its ability to foster human goods in general. That trust is reasonable just in case it survives conscientious reflection. In this way of looking at faith, it is part of a set of beliefs, emotions, and acts that are all reasonable if they satisfy the following condition: they arise from and are maintained in an epistemically conscientious way, and they survive conscientious reflection.

I said above that ultimately, our only test that we have reached the truth in any given case is survival of conscientious reflection. We reflect upon all our conscious states, evaluate them according to norms that we have adopted as the result of previous reflections, and adjust our beliefs in an attempt to make those states survive future conscientious self-reflection, given that we expect there will be changes to the self with new experience, new memories, changes in what we feel and what we trust, and changes in other beliefs. When a conflict arises within our beliefs, we reflect upon both beliefs in conjunction with our other beliefs, and retain the one that we conscientiously judge is most likely to survive conscientious self-reflection in the future. A conscientious person does the same thing when one of the beliefs is a belief of faith. That is all a reasonable person can do.

References

Alston, William P. 1986. "Epistemic Circularity," *Philosophy and Phenomenological Research* 47, 1–30.

Alston, William P. 2005. *Beyond Justification: Dimensions of Epistemic Evaluation*. Ithaca, NY: Cornell University Press.

Anscombe, G. E. M. 1979. "What is It to Believe Someone?" in *Rationality and Religious Belief*, edited by C. F. Delaney. Notre Dame, IN: University of Notre Dame Press, 141–51.

Bacon, Francis. 2000 [1605]. *The Advancement of Learning*, edited by Michael Kiernan. New York: Oxford University Press.

Foley, Richard. 2001. *Intellectual Trust in Oneself and Others*. New York: Cambridge University Press.

Grice, Paul. 1989, *Studies in the Way of Words*, Harvard University Press,

Lehrer, Keith. 1997. *Self-Trust: A Study of Reason, Trust, and Autonomy*. Oxford: Clarendon Press.

Locke, John. 1975 [1689] *An Essay Concerning Human Understanding.* Edited by P. Nidditch. Oxford: Clarendon Press.

McMyler, Benjamin. 2011. *Testimony, Trust, and Authority*. NY: Oxford University Press.

Menssen, Sandra and Thomas D. Sullivan. 2002. "The Existence of God and the Existence of Homer: Rethinking Theism and Revelatory Claims," *Faith and Philosophy* 29, 331–47.

Moran, Richard. 2005. "Getting Told and Being Believed," *Philosophers' Imprint* 5:5, 1–29; reprinted in *The Epistemology of Testimony*, edited by Jennifer Lackey and Ernest Sosa, NY: Oxford University Press.

Plantinga, Alvin. 1983. "Reason and Belief in God," in *Faith and Rationality*, edited by Alvin Plantinga and Nicholas Wolterstorff. Notre Dame, IN: University of Notre Dame Press, 16–93.

Plantinga, Alvin. 2000. *Warranted Christian Belief,* New York: Oxford University Press.

Zagzebski, Linda. 2011. "First Person and Third Person Epistemic Reasons and Religious Epistemology, *European Journal of Philosophy of Religion* 3(2), 285–304.

Zagzebski, Linda. 2012a. *Epistemic Authority: A Theory of Trust, Authority, and Autonomy in Belief*. NY: Oxford University Press.

Zagzebski, Linda. 2012b. "Powers and Reasons," in *Powers and Capacities in Philosophy*, edited by John Greco and Ruth Groff. New York: Routledge, 270–82.

Zagzebski, Linda. 2014."First Person and Third Person Reasons and the Regress Problem," in *Ad Infinitum: New Essays on Epistemological Infinitism*, edited by John Turri and Peter Klein, New York: Oxford University Press, 2014, pp. 243–55.

Zagzebski, Linda. 2015. "Faith and Believing a Person," in *Christian Philosophy of Religion: Essays in honor of Stephen Davis*, edited by Colin Ruloff, Notre Dame, IN: University of Notre Dame Press.

CHAPTER 22

EXPERIMENTAL PHILOSOPHY

RON MALLON

1. Introduction

"Experimental philosophy" labels an approximately 15-year-old movement of philosophers and psychologists using empirical methods—especially those of social psychologists—in the hopes of illuminating philosophical questions.[1],[2] Experimental philosophy has been controversial in part because of the general methodological questions it raises as well as for the myriad specific claims surrounding specific experimental findings. My aim here is to characterize a few central metaphilosophical strands of discussion in experimental philosophy, and suggest a new empirical possibility—intra-individual diversity—that requires the rethinking of both old and emerging debates.

Because experimental philosophy is a movement characterized by endorsement of the use of experimental methods, there is no deep ideological coherence among experimental philosophers as a whole, save the simple assertion that experimental methods can (at least in principle) reveal facts that bear on at least some philosophical theses—a view shared with a much broader range of philosophers and most critics of experimental philosophy. Nonetheless, many experimental philosophers endorse some variety of an older, *Naturalists' Challenge* to the traditional use of intuitions in philosophical theorizing, one that holds that the philosophical use of intuitions must be reformed or abandoned in light of emerging evidence from the sciences (see, e.g. Cummins 1998, Goldman and Pust 1998, Kornblith 2007, Stich 1990).[3]

[1] Thanks to Joshua Alexander, Thomas Nadelhoffer, Eddy Nahmias, Shaun Nichols, Ángel Pinillos, Jonathan Weinberg, and two anonymous referees for help with or comments on earlier drafts of this chapter.

[2] While "experimental philosophy" had other, cognate uses in the past (for example, as the title of a reading group Patricia Churchland ran at UC- San Diego in the 1980s), this contemporary use of "experimental philosophy" may have been coined by Brandon Towl. around 2001. It likely became entrenched because of the eponymous blog founded by Thomas Nadelhoffer in 2004.

[3] "Naturalism" in a *methodological* sense of naturalism, one that (at least) takes the methods of the natural sciences to be a central epistemic route to true beliefs. The Naturalists' Challenge is consistent with *metaphysical* non-naturalism about the properties in question (see Prinz 2007).

And we can find greater ideological coherence if we look to subgroups of experimental philosophers. So-called *negative* or *negative program* experimental philosophers view experimental work as revealing that the use of philosophical intuitions as evidence for the truth of a philosophical claim is bankrupt. Other, *positive* or *positive program* philosophers see experimental methods as offering the possibility to continue and improve upon traditional uses of intuitions (e.g. Glasgow 2008, Nadelhoffer 2006a, 2006b; Murray and Nahmias 2014, Nahmias and Thompson 2014).[4] And still others hope to escape this dialectic altogether, for example by viewing the primary contribution of experimental work as the empirical investigation of facts about human nature or human understanding of philosophical interest (e.g. Knobe and Nichols 2008), or by illuminating scientific concepts of import to philosophers of science (e.g. Stotz et al. 2004).

A great deal of interesting and important work is being undertaken in all of these areas. In what follows, I will focus broadly upon experimental philosophy as an extension of the Naturalists' Challenge to the use of intuitions in philosophy, and to traditional or "armchair" responses to this challenge. I begin in section 2.1 by discussing the role and nature of intuitions, and two sorts of experimental philosophical challenge to them: *the challenge from irrelevant determination* and *the challenge from diversity*. In section 3.1, I argue that the strongest experimental philosophical challenge uses the challenge from diversity to undermine the reliability of intuitions as evidence for the truth of philosophical claims. In section 4.1, I consider critical responses to the challenge from diversity and I explore how those challenges have played out in experimental philosophical accounts of reference. One of these responses, what I call the "Multiple Signals" response, suggests that diverse responses to experimental vignettes may be a consequence of verbal disagreement, that is, of different conceptual competences used to construe the same vignette. In section 5.1, I explore the possibility that this same phenomena may occur within individuals, and consider its implications for the Naturalists' Challenge.

2. Intuitions and Two Kinds of Experimentalist Challenge

While there is a wide range of experimental philosophical findings that bear on philosophical questions, experimental philosophy has seemed especially provocative because of its role in extending the Naturalists' Challenge, calling into question the role intuitions play in philosophy. In this section, I explore the role intuitions play in philosophical discussion, and what intuitions are. I then focus on two different sorts of experimental challenges: *the challenge from irrelevant determination* and *the challenge from diversity in intuitions*.

[4] See Alexander, Mallon, Weinberg 2010a. The distinction between the negative and positive programs, as presented here, is coextensive with Alexander and Weinberg's (2007) distinction between the *restrictionist* view and the *proper foundation* view, but cross-cuts the useful distinction drawn by Nadelhoffer and Nahmias (2007) between "experimental analysis" (roughly, *what* do people think) and "experimental descriptivism" (*how*—by what processes or mechanisms—do they think it?).

2.1 Intuitions and Philosophical Theorizing

While there is no real consensus as to just exactly what intuitions are, ostensibly the term picks out the relatively quick responses to actual and hypothetical situations that are characteristically elicited in philosophical discourse and that often play a role in constraining the discourse. Intuitions are typically elicited by describing (sometimes mundane and sometimes outlandish) hypothetical situations, though they may also be elicited by actual situations.

Perhaps the most famous contemporary example of this practice is Edmund Gettier's pair of compelling counterexamples to understanding knowledge as justified true belief (1963)—examples widely taken to show that the justified true belief analysis of knowledge is insufficient. But other examples are not hard to find; indeed, philosophy seems rife with them. We have, to name a few, intuitions about the morality of diverting or blocking runaway trolleys, intuitions about knowledge in fake barn country, intuitions about a Swampman, about the possibility of phenomenal "zombies," about the neuroscientist Mary who was raised in a black and white environment, about whether a Chinese room can think, and about whether the Chinese nation (appropriately organized) could be conscious.[5]

These well-known cases and others like them are our *target cases,* ostensively fixing the sort of states, methods, and thought experiments that experimental philosophers and their critics have in mind in discussing intuitions.

One common way of understanding such target cases is that they involve a *Method of Cases* in which intuitive, "armchair" responses to actual and possible states of affairs serve as evidence for the truth of one or another philosophical theory. In talking of "armchair" responses, I mean responses that draw upon one's existing beliefs and competencies, and perhaps those of one's interlocutors, but do not involve any systematic empirical investigation. While such responses may be a priori, armchair responses can be (and no doubt often are) guided by one's (experientially acquired) beliefs.

2.2 What are Intuitions?

While there is little consensus about the exact psychological or epistemic character of intuitions, they are widely treated as a variety of propositional attitude. Some (e.g. Sosa 2007, Bealer 1998) treat this attitude as a *seeming* or a *temptation* to believe *p*, while others have it that the attitude is simply a *judgment* or *belief* that *p* (e.g. Sinnott-Armstrong 2008, McMahan 2000).

With this distinction in mind, it is tempting to think that, other things equal, "seeming"-type intuitions cause "judgment"-type intuitions: when it intuitively seems to S that *p*, then, other things being equal, S comes to believe or judge that *p*. However, it's not clear that this model is correct, and other models could as easily be true (e.g. subjects initially *believe*

[5] For trolleys: see Thomson 1985; fake barns: Goldman 1976; Swampman: Davidson 1987; zombies: Chalmers 1996; Mary: Jackson 1982; the Chinese room: Searle 1980; the Chinese nation: Block 1978.

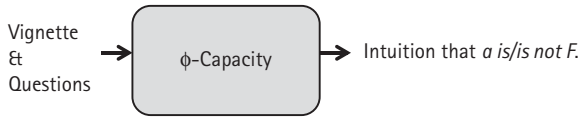


FIGURE 22.1 Simple model of a thought experiment

p, but then, if necessary, withdraw belief). In any case, because "seeming"-type intuitions are not themselves all-out beliefs or judgments that *p*, such temptations are consistent with S's also having other (either "seeming"-type or "judgment"-type) intuitions that not-*p*.

Many philosophers also understand intuitions as the "spontaneous" result of non-inferential, non-perceptual mental processes in which intuitions that *p* simply emerge among one's conscious attitudes without conscious reasoning to them. Because they are spontaneous, such states seem to arrive in advance of their warrant, and so raise the question as to whether the possession of an intuition by itself can be evidence for the truth of a philosophical claim.

To this, we can add a bit more. Call whatever guides our intuitive responses to thought experiments our *Φ-capacity*. A *simple model of a thought experiment*, then, has it that a vignette or situation description, along with questions about it, is input to some capacity we have for producing intuitions, eliciting an intuition about the entities described in the vignette (see Figure 22.1). As this model suggests, philosophical intuitions (including those we have discussed) are often categorizations—judgments that some thing falls or does not fall under some concept. We judge that some mental state is not an instance of *knowledge*, or that some action is not *morally permissible*, or that some state of affairs is *possible*. Noting this, one common view among contemporary philosophers has been that our Φ-capacity is primarily a *conceptual competence* with one or more concepts that figure in the propositional object of our intuition (e.g. Ludwig 2007, 2010). Others, however, have insisted our Φ-capacity is actually is a much broader competence (including, perhaps, all the information the subject has at their disposal). Michael Devitt, for instance, characterizes intuitions as "empirical theory-laden central processor responses to phenomena" (Devitt 2006, pp. 103ff). And Timothy Williamson writes, "so-called intuitions are simply judgments (or dispositions to judgment); neither their content nor the cognitive basis on which they are made need be distinctively philosophical" (Williamson 2007, p. 3).

Throughout this chapter (except where explicitly noted) I remain uncommitted on the exact character of intuitions. I use "intuition" both to pick out the sorts of seemings or judgments involved in our target cases but also, more loosely, I speak of the behavioral manifestations of those seemings or judgments produced in response to thought experiments and experimental surveys as subjects' intuitions.[6]

[6] Though see Kauppinen 2007, p. 107; Cullen 2010.

2.3 The Challenge from Irrelevant Determination

Philosophical naturalists have long expressed skepticism about whether intuitions about cases appropriately constrain philosophical theorizing (e.g. Stich 1990; Cummins 1998), and experimental philosophers have extended these challenges by turning naturalist conjectures and existing research methods into empirical evidence suggesting epistemic failings for the armchair.

One important class of failings suggests that intuitions vary in response to influences that are irrelevant to the domain. In these cases, "intuitions" are typically applications of philosophically important concepts (*knowledge, moral permissibility*, etc.) to particular instances. Examples of results of this form include:

Order effects: Intuitions can vary as the result of the order in which experimental vignettes are considered. So, for example, Stacey Swain, Joshua Alexander, and Jonathan Weinberg (Swain et al. 2008) presented a number of vignettes to subjects, including a version of Keith Lehrer's (1990) Truetemp Case, and they asked if each was a case of knowledge. They found that:

> Compared to subjects who receive the Truetemp Case first, subjects first presented with a clear case of knowledge are less willing to attribute knowledge in the Truetemp Case, and subjects first presented with a clear case of non-knowledge are more willing to attribute knowledge in the Truetemp Case. (138)[7]

More recently, Eric Schwitzgebel and Fiery Cushman offer evidence of order effects on moral judgments made by professional philosophers (2012).

Because the order in which vignettes are presented makes no difference to what situations the vignettes represent, the order is presumably irrelevant to the truth of claims about knowledge in the cases. Thus effects of order on judgments about the cases amount to irrelevant determination.

Framing effects: Intuitions vary as the result of non-representational differences in how vignettes are posed. (We will see the most famous example of this—Amos Tversky and Daniel Kahneman's "Asian Disease" case, in section 5.1.1) Framing effects concern vignettes that are extensionally equivalent or describe the same set of possible worlds, but they nonetheless can elicit different intuitions leading to apparently contradictory responses.[8] Since the correct judgment presumably depends exclusively on the facts of the situation described, such non-representational differences are irrelevant to the truth. Walter Sinnott-Armstrong (2008) has argued that such effects suggest that we ought not to rely on intuition in moral philosophy.

Functional biases: Intuitions are biased in a range of ways that are plausibly products of evolutionary adaptations. For example, Petrinovich and O'Neill (1993) found that responses to trolley cases were consistent with a subject's tracking the inclusive fitness of

[7] Interestingly, Jen Wright (2010, 2013) has argued that subjects' confidence in their judgments is inversely correlated with the instability of those judgments in response to order effects. This suggests the possibility of an introspectively accessible way of ruling out such instability.

[8] That is to say that the vignettes and questions, taken as a whole, describe the same situations or possible worlds. The components of the vignettes are not equivalent in this way.

various possible individuals. More recently, Joshua Greene and colleagues (Greene et al. 2001, Greene 2007) have suggested that emotional responses to trolley cases are governed by fast, heuristic appraisals adapted to the circumstances of our evolutionary past. Crucially, however, we lack a reason to think that mechanisms adapted to our survival and reproduction in our evolutionary past should converge upon truth in the moral domain, and without it, it seems that functional biases determine moral judgment in ways that are irrelevant to the truth (Greene 2003, 2007; Singer 2005, Joyce 2006). Richard Joyce puts it thus: "We should reject or modify any theory that would render us epistemic slaves to the baby-bearing capacity of our ancestors" (2006, p. 219).

These are just a few of many putative examples of potentially irrelevant determination of philosophical judgments. Others include:

- Effects of disgust (Schnall et al. 2008).
- Effects of the font in which a vignette was presented (Weinberg, Alexander, Gonnerman, and Reuter 2012).
- Effects of the abstractness or concreteness of an event description (Nichols and Knobe 2007).
- Effects of the temporal distance of an event (Weigel 2011).
- Effects of circumstances being actual or merely possible (Roskies and Nichols 2008).

One way that irrelevant determination might threaten the evidential value of intuition is by *debunking* them (Mason 2011; Mallon and Doris 2013). Intuitions are debunked to the extent that they are causally explained by appeal to mechanisms or processes that we believe are not epistemically good indicators of the domain (where "epistemically good" can be filled in in a number of ways—e.g. insufficiently reliable). These explanations, in turn, causally exclude (or partially exclude) explanations that explain the intuitive content *p* by appeal to the fact that *p*.

In the case of functional biases, this suggestion of epistemic insufficiency is strengthened to the extent we have reason to believe that the mechanisms or processes in question function to track facts that, while evolutionarily important, are not plausibly correlated with the facts in the target domain.

2.4 The Challenge from Diversity in Intuitions

A different sort of naturalist challenge emerges from the possibility that individual persons or groups of people may have divergent intuitions about target cases.[9] In recent years, a range of evidence suggesting such variation has emerged. For example:

Cultural diversity: Philosophical intuitions can seem to differ according to the culture of the intuiter.

[9] Not everyone understands it as very different. Some interpret evidence of cultural diversity as continuous with the other challenges of irrelevant determination that have been suggested, e.g. Weinberg et al. 2010, Feltz and Cokely 2012.

Inspired by cross-cultural work on cognitive differences by Richard Nisbett and colleagues (for a summary, see Nisbett 2003), Jonathan Weinberg, Shaun Nichols, and Stephen Stich (2001) set out to find out whether epistemological intuitions might similarly vary, and found preliminary evidence of systematic differences in response to epistemological thought experiments (including TrueTemp and Gettier cases) between cultural East Asians and Westerners among New Jersey undergraduates, and also among groups divided by "high" and "low" socioeconomic status. This study has led to a number of more recent studies that have called into question the initial findings. Among these, some have been unable to replicate Weinberg et al.'s results (Nagel 2012, Nagel et al. 2013, Seyedsayamdost (2015)). Some have explained Weinberg et al.'s results as an unintended artifact of the experimental materials or situation (e.g. Nagel 2012, Cullen 2010, Turri 2013). And John Turri (2013) argues that the effect goes away for Gettier cases when you structure the experimental probes into stages.

Also inspired by Nisbett, others have sought diversity in other domains. Edouard Machery, Nichols, Stich, and I (Machery et al. 2004) found preliminary evidence of cross-cultural differences in intuitions about the reference of proper names. We return to consider this case in more depth in section 4.1.3.

Other sources of diversity in response to vignettes include:

> *Personality traits:* Intuitions vary according to personality traits and other individual differences. For instance, Adam Feltz and Edward Cokely (2008) found that judgments about moral objectivism varied according to possession of the personality trait *openness to experience*. And Cokely and Feltz have also found that possession of the trait of *extraversion* predicts responses on questions regarding free will and moral responsibility (Cokely and Feltz 2009).
>
> *Conceptual Diversity:* Intuitions may exhibit diversity because they are manifestations of distinct concepts. For instance, Shaun Nichols and Joseph Ulatowski (2007) have suggested that judgments about whether foreseen but unintended side-effects are judged to be intentional vary because they are guided by different implicit concepts or conceptions.

While most surveys exhibit some diversity in responses, these studies go beyond that to indicate *systematic* diversity. Such systematic diversity cannot be dismissed as mere statistical noise caused by the host of distorting factors that attend any study. Rather they demand an explanation stemming from an independent variable. Data suggesting differences in philosophical intuitions between members of East Asian and Western cultures, for instance, requires an explanation that stems from the differences in the subject populations.

If philosophical intuitions are not widely shared, it calls into question both their dialectical relevance and their truth. It calls into question their dialectical relevance since an unshared assumption (be that an implicit theory or competence giving rise to the intuitions, or the intuitions themselves) is often a poor foundation for further persuasive argument. And it calls into question their truth since disagreement among individual or groups of epistemically parallel agents is prima facie evidence of possible error on the part of at least some of the agents. Here, I focus on truth.

Evidence of diversity of intuitions can be addressed in at least three ways:

1. One option is to regard diversity as evidence that undermines the reliability of intuition in general. The failure of epistemically parallel agents to agree suggests that intuition is a poor source of evidence of some common subject matter, and so perhaps it should simply be abandoned—a possibility amenable to negative experimental philosophers. Call this interpretation of disagreement *All Noise*.
2. A second, more conservative and more common response to philosophical disagreement is to regard diversity as evidence that one or more parties has made a mistake or has failed to properly attend to the subject matter. Ideally, this interpretation is accompanied by an account of the nature of the *signal* to be attended to, and the *noise* that gives rise to the mistake. Call this the *Noise and Signal* response.
3. A third interpretation suggests that diversity is the result of multiple different, but equally legitimate "signals" of philosophical interest. Call this interpretation *Multiple Signals*.

Because the latter two interpretations apparently have less sweeping consequences for philosophical methodology, they amount to two routes for defending philosophical intuitions from versions of the Naturalists' Challenge rooted in diversity in intuitions, and we return to these interpretations in section 4.1. I now turn to consider the reliability of intuitions.

3. Intuitions as Reliable Evidence

Why do intuitions constrain philosophical theorizing? Perhaps the most obvious way to understand the method exhibited in the target cases is the one we have already suggested: the Method of Cases takes the intuitions to be evidence for the truth of philosophical claims. This seems to presuppose something like:

Reliable: Intuitions are reliable guides to the truth of philosophical claims.

If Reliable is true, it rationalizes philosophical use of the Method of Cases in target cases. If it is false, it undermines this practice. This makes it seem just the sort of principle to be at the heart of the dispute between experimental philosophers and defenders of the Method of Cases.

However, recently two alternate accounts of the central philosophical issue threaten to sideline *Reliable* as the crux of the debate.

3.1 Looming Skepticism

One response to the Naturalists' Challenge is that intuitions are so ordinary and widespread as a source of belief that to doubt them en masse—to doubt Reliable—is to invite a radical skepticism that would be at least as devastating to the naturalists claims of scientific knowledge as to defenders of armchair intuition (Bealer 1998, Sosa 2007).

Some have developed this idea by pressing a comparison of intuition and perception. Ordinary perception offers a pretheoretic and non-inferential source of belief. And like the challenges to intuition, psychologists have shown perception to be susceptible to a range of systematic mistakes and distortions. This comparison suggests an answer to one part of the Naturalists' Challenge. The answer begins with the thought that, on nearly all accounts, perception is a reliable and legitimate source of knowledge despite its well-documented failings in the vision lab. But as with perception, so with intuition. The mere fact that, say, framing effects or order effects can be shown to affect intuitive judgments is not, by itself, enough to disqualify intuitive judgment as a source of knowledge. In using empirical evidence to challenge the reliability of intuition, the naturalist threatens perception as well and invites skepticism (Sosa 2007).

Swain and colleagues respond to this general line of argument in two, interconnected ways. First, they have argued that in the case of perceptual knowledge we have a good folk theory of perception that specifies, for example, good observation conditions, but we have nothing of the sort for intuitions (Swain et al. 2008, pp. 147ff; cf. Weinberg 2007, p. 326). At the very least, experimental work is needed to show us how reliable a guide to truth our intuitions are. They write:

> At this time, we don't know what is the parallel for intuition of making sure that the light is on; that is, we do not know which are the circumstances that render intuition reliable or unreliable. With perception, by contrast, we are aware of the dimensions of variance and how to compensate for them. We know to turn our heads toward the speaker if we cannot hear well, or to squint if we are trying to read a distant road sign, or to cleanse the palate before evaluating a fine wine. What our research indicates is that we do not have analogous knowledge relating to our practice of relying on intuitions.
>
> (Swain et al. 2008, 148)

Swain and colleagues overstate their case here, and in both directions. Our folk theory of perception doesn't tell us how to avoid the visual illusions that scientists can produce in us. And we plausibly have a clearer background theory than they suggest, allowing us to discern which intuitions count and which do not. We have already noted, for example, our sense in the moral domain that certain differences that may be evolutionarily important may not be morally relevant. In fact, the methodology of thought experiments is shot through with such assumptions about relevance. Just as with empirical experiments, philosophical thought experiments are best designed precisely to get at particular factors of interest. The elaborate design of thought experiments about trolleys, for example, is an attempt to elicit precisely the intuitions that matter while recognizing that there are many factors that do not. Similarly, theoretical constructions like Rawls's "original position" might be thought of as cultural prosthetics or measuring instruments designed to elicit exactly the intuitions relevant to some philosophical question, but they are produced through constructive processes guided (perhaps imperfectly) by folk conceptions of what is relevant to the moral domain (Rawls 1971). In addition, empirical evidence may also reveal that we already have *some* means of discerning good and bad intuitions from the armchair (Wright 2010, 2012; Nagel 2012).

A second response by Weinberg (writing alone) goes beyond pressing disanalogies with perception that might bear on Reliable to the more radical and interesting claim

that experimental work ought not to be understood as challenging Reliable, even if we understand "being reliable" as requiring great success in indicating the truth in a domain (Weinberg 2007). Weinberg stops short of questioning what he calls "practically infallible" mechanisms, but he does assert the epistemic insufficiency of even very accurate intuitions, suggesting that even a very small error rate for intuition is too high in the absence of an epistemic virtue that he calls "hopefulness" (2007, 324–325). Hopefulness is our capacity to "detect and correct" the errors, something he thinks that vision has and intuition does not.[10] While this discussion is provocative and fruitful, I leave it aside for now. In conceding reliability (even understood as entailing high accuracy in judgment), the focus on hopefulness concedes a great deal to the defender of intuition. If intuition is a very reliable source of knowledge, that by itself may be enough to vindicate the Method of Cases as philosophical methodology (even if a hopeful methodology would be better). In any case, a way of answering the challenge of looming skepticism that didn't concede reliability would surely be preferable, especially since reliability is at the heart of the way many others have understood the Naturalists' Challenge (e.g. Sosa 2007; Sinnott-Armstrong 2008; Goldman 2010).

Is there such a way? Return again to the analogy with perception. Perception is subject to a range of distortions that are revealed in visual illusions, and that these distortions seem analogous to some of the distortions revealed by work on intuitions. The result is that the connection of such biases with undermining reliability is far from clear. Defenders of functional biases might make a better case, *if* they can show not only that a mechanism is *biased* by evolutionary considerations, but also that it *primarily tracks* some evolutionarily important property that has no plausible correlation with the property of philosophical interest.[11] However, nothing like this has been shown for any case of interest.

However, as we have seen, some experimental philosophy emphasizes, not the challenge from irrelevant determination, but the challenge from diversity. The challenge posed by diverse answers to vignettes can be read as exhibiting further the determination of responses by irrelevant influences, but a more powerful way to read the challenge begins with the fact that agents (or groups of agents) who are apparently epistemically on par seem to disagree about crucial philosophical cases. This suggests the possibility for undermining Reliable in a straightforward way: contradictory responses cannot all be right. While this view may have other undesirable consequences, it does not provoke looming skepticism since it begins with evidence of diversity in a domain, and its skeptical conclusions (at least initially) concern only that domain.

So there is a split in the Naturalists' Challenge. Those who are most concerned with the influence of irrelevant factors on intuition need a way to answer the charge of looming skepticism. Perhaps Weinberg's emphasis on "hopefulness" will do. On the other hand, those whose concern grows out of disagreement have a prima facie argument for unreliableness of many intuitions (even without knowing which ones).

[10] For a response to Weinberg, see Ichikawa 2011.

[11] Haidt (2001), Greene and Haidt (2002), and Greene (2007) can be read as implying this in the moral domain. Cf. Mallon and Doris (2013) for further discussion.

3.2 Are Intuitions Evidence?

Thus far, I have assumed, along with many experimental philosophers as well as many of their critics, that philosophers use the Method of Cases. We have assumed what Herman Cappelen has recently called *Centrality*, that "Contemporary analytic philosophers rely on intuitions as evidence (or as a source of evidence) for philosophical theories" (2012, p. 3). But recently, Centrality itself has come into question with some critics charging that existing philosophical practice does not rely on intuitions as evidence. As such, it is not in need of reform even if naturalist doubts about intuitions are vindicated.

Williamson has argued that philosophy has no distinctive method (Williamson 2007), and so naturalist critics of armchair methodology have no target. Experimental philosophers, he writes, have failed "to identify any distinctive philosophical method to be transformed or overturned by their revolution" (2013, p. 472), and he suggests that experimental philosophers are critically engaging a now abandoned position:

> Experimental philosophers did not invent the idea of "philosophical intuition". It belonged to the ideology of one faction of the *ancien régime*. Against that faction, their use of it was dialectically legitimate. For constructive purposes, however, it has outlived its utility. The psychological and sociological study of philosophy will make more progress once it ceases to work within a framework of obsolescent epistemology.
>
> (474)

Williamson's critique of experimental philosophy can be puzzling for what he lays at the feet of experimental philosophers—a concern with the role of intuitions in philosophical method—belongs at the feet of a much broader community of philosophers: not only those who offer the Naturalists' Challenge, but also many of those those who defend armchair intuition against it. To read him, it is easy to think that no one actually takes intuitions or the Method of Cases seriously. But looking around, there are numerous apparent counterexamples among both critics and proponents of the Naturalists' Challenge. And Williamson's argument cuts at least as deeply against many of the targets of experimental philosophy as it does against its practitioners.[12] Declaring a focus on the role of philosophical intuition in philosophical method, "obsolescent epistemology" has the odd effect of suggesting that the recommendations of the most extreme naturalist critics of intuition (the complete abandonment of intuition as a source of evidence) have already been widely conceded, when they clearly have not.[13]

Max Deutsch (2009, 2010) and Cappelen also think Centrality is false, but, in contrast to Williamson, they acknowledge that a great many philosophers take it to be true, and as such, denying it needs to be argued for. They each undertake to do so by careful reconstruction of target cases in which they hope to reveal a very different method at work. At the same time, they agree with Williamson that philosophical reliance on armchair intuition as evidence would be problematic and that existing philosophical method involves

[12] For example, George Bealer (1998), Michael Devitt (2011), Alvin Goldman (2010), Frank Jackson (1998), Antti Kauppinen (2007), Kirk Ludwig (2007, 2010), Jennifer Nagel (2012), and Ernest Sosa (2007) all defend (some variety of) philosophical intuition against naturalist and experimentalist critics.

[13] For further critical discussion of Williamson, see Alexander 2012; Weinberg 2009.

Table 22.1 Simple Model of a Thought Experiment.

		Methodological Reform is…	
		Not Needed	Needed
Armchair Intuitions are…	Not a proper source of evidence for philosophical inquiry.	**Negative Armchair Philosophers:** e.g. Cappelen, Deutsch, Williamson	**Negative Experimental Philosophers:** e.g. Alexander, Machery, Mallon, Nichols, Stich, Weinberg
	A proper source of evidence for philosophical inquiry.	**Positive Armchair Philosophers:** e.g. Bealer, Jackson, Kauppinen, Ludwig, Nagel[33]	**Positive Experimental Philosophers:** e.g. Glasgow, Nadelhoffer, Nahmias

[33] Nagel defends armchair intuition experimentally (e.g. Nagel 2012, Nagel et al. 2013).

little or no such reliance. Cappelen offers an extended defense of the view that, rather than indicating an appeal to evidence, "intuition" talk plays a number of roles and often simply marks assertions as having a certain dialectical role in an argument.[14]

In denying Centrality, Williamson, Deutsch, and Cappelen open up space for combining a defense of existing philosophical armchair practices with acknowledgement of critics' charges that appeals to intuition as evidence are defective. In effect, they have recreated the rift over the proper role for intuitive judgments that divides negative and positive experimental philosophers among defenders of "armchair" methods. As such, we might call the resulting positions *negative armchair philosophy* (see Table 22.1).

If the negative armchair philosophers are right, and Centrality is false, then much of the methodological anxiety surrounding experimental philosophical work by practitioners and opponents alike is irrelevant, and the primary contribution of the Naturalists' Challenge to philosophy will have been to make clearer the actual role of the target cases in philosophical work. Nonetheless, in what follows, I ignore these challenges to Centrality, and the challenge that the denial of Centrality poses to the importance of Reliable. As Deutsch and Cappelen recognize, an evidential role for intuitions is a widely shared methodological assumption among philosophers. Embracing its denial involves rereading many philosophers in ways that they themselves do not seem to endorse and, in some cases, seem to not endorse. This offers a strong (albeit prima facie) reason to think it is mistaken at least to some significant extent.[15] And to whatever extent it is mistaken, understanding the use of intuitions as evidence is crucial to assessing philosophical practice.

[14] Cappelen 2012, ch. 2. Cf. Dorr 2010.

[15] The reasons to think Centrality is true in some of these cases parallel the reason to think that it is false in the case of, say, Cappelen's own current work, given his repudiation of it.

3.3 Reliability

Accepting that philosophers employ the Method of Cases, we still need some account as to why anyone should take it seriously. The answer we have been sketching is that philosophers who use it assume Reliable. But then *why* should Reliable be true? Understanding this requires some account of what intuitions are evidence for, and, unfortunately, there is some disagreement about this question as well. Here I distinguish three accounts, the second a version of the first.

The first account holds that a person's (or perhaps group of persons') having an intuition that *p* in favorable conditions is evidence for the truth of *p*. (Though the label is not perfect) I'll call this the *disquotational account* of intuitive knowledge. The disquotational account will be true if the mechanism underlying the intuitions—what we labeled the Φ-capacity in our simple model (see Figure 22.1 and section 2.2)—reliably indicates the facts that the intuition is about under favorable circumstances.

A second view presupposes (and is a version of) the first. It begins with the suggestion that the primary function of intuitions in philosophy is to give evidence about *extra-mental facts*, for example, about some extra-mental property, universal, or natural kind.[16] We can further distinguish two different versions of this extramentalist view.

One Platonist possibility is that our Φ-capacity allows us to *directly* apprehend some fact that *p*, where *p* is some universal, abstract object, or general principle. For example, perhaps we can directly apprehend truths about freedom or justice or knowledge or about logical or mathematic facts like modus ponens. Perhaps we simply directly intuit contents like:

> Ought implies can.

Or:

> *P* and *If P then Q* entails *Q*.

A second possibility is that we gain *indirect* knowledge of such very general extramental facts by inferring from the content of intuitions about particular cases. Consider the Gettier intuition. Since Smith has justified true belief and we intuit that,

> *Smith does not have knowledge*

It follows that knowledge is not justified true belief. Thus construed, the intuition that *p* is a premise in the inference to a conclusion about the nature of knowledge.

Such extramentalism seems to fit closely with how intuitions are actually used in some philosophical practice—they are treated as giving us evidence about the nature of things in the world. But both versions of extramentalism presuppose (and are versions of) the disquotational view, since both versions rely on our intuition that *p* indicating *p*. This is perhaps obvious for the direct view, but it is also true of the indirect view. The ability to draw true inferences about, for example, knowledge from intuitions whose propositional object includes the concept *knowledge* depends upon the propositional objects of these intuitions being true. It is because we take "Smith does not have knowledge" to be true that we think it allows us to infer something about knowledge itself.

[16] Goldman and Pust (1998) introduce this distinction with this terminology.

Extramentalism depends upon the reliability of the Φ-capacity in indicating the truth in the domain in question. But what is the Φ-capacity such that it should allow armchair intuitions that *p* to be reliable indicators that *p*? Unfortunately, there is no broad consensus as to how to answer this question, no "accepted explanation of the hope-for correlation between our having an intuition that P and its being true that P" (Williamson 2007, 215).

Many philosophers, perhaps worried that no naturalistically acceptable account of intuitive knowledge of extramental features of the world is forthcoming have instead sought to develop the suggestion that intuitions serve in the first case as evidence about, or a manifestation of, *mental* entities—for example, evidence for the character of the theories, concepts, conceptions, or other psychological competences that give rise to the intuition (e.g. Goldman and Pust 1998; Alexander et al. 2010; Goldman 2010; Jackson 1998). For the mentalist, intuitions are, in the first case, reliable indicators of our Φ-capacity.

For instance, in the Gettier case, the intuition that *Smith does not have knowledge* is used to reconstruct the concept *knowledge* (or the psychological mechanism or vehicle that subserves or expresses this concept). I will call the Φ-capacity that is the object of this *mentalist* project the conceptual competence (though I remain neutral on its exact character—whether it is an abstract entity or a psychological mechanism that produces or expresses that entity). We should add that it remains open to the mentalist to hold that knowledge about our psychological competences provides some knowledge about the world itself, namely, what things would have to be like to fall under the concept (Jackson 1998).[17]

Alvin Goldman and Joel Pust argue for the plausibility of a mentalist approach to philosophical method by emphasizing the plausibility of the view that the character of a conceptual competence F might be reliably indicated by our F-involving judgments (188ff).[18] To this plausibility, we can add that in contrast to extramentalism's endorsement of the content of intuitions, the inference from "I intuit that *a* is *F*" to the conclusion that "My *F*-concept is such and so" does not require taking the proposition *p* (here, the content a *is* F) to be true. So, in contrast to extramentalism, mentalism need

[17] The extramentalist has a complementary advantage: extramentalism leaves open the possibility that intuitions that *p* might be evidence for extramental facts without any need to successfully reconstruct the conceptual competence involved in the production of the intuition.

[18] This overstates the case, for the competences we draw inferences about are often not narrowly concerned with the particular word meanings or concepts that figure in the intuitions. For example, we may infer from the intuition that,

It would be impermissible to push the man off the footbridge to die in front of a trolley

to the judgment that,

Utilitarianism is inadequate as an account of folk morality

though the concepts in the latter proposition are (mostly) not present in the former. To choose another example that will be relevant later in the chapter (see section 4):

intuitions about the application of names or natural kind terms are used to draw inferences about the folk understanding of the reference relation for kinds of terms, though the relevant intuitions do not typically contain or express concepts like *reference*.

make no commitment that intuitions that *p* are reliable indicators of or "correlated with" *p*.

So far, these different accounts of philosophical practice have very different commitments regarding the reliability of intuition. The disquotational account and the extramentalist account stand or fall with the reliability of the capacity which gives rise to our intuitions in indicating facts in the domain in question (where the extramentalist understands these as non-mental, non-psychological facts about the world). The mentalist position requires only the reliability of intuitions in indicating facts about the conceptual competence itself. It seems that while extramentalism promises knowledge of a far greater and more important range of phenomena, it also is committed to capacities for knowledge that are themselves poorly specified and understood. As such, it might seem to be much more vulnerable to naturalist attacks on the reliability of intuition. In contrast, mentalism seems to promise a much more epistemically palatable account of philosophical practice, albeit one with more modest aims. And so it might seem (as it has seemed to naturalists like Goldman and Pust) that mentalist construals of philosophical method are less vulnerable to experimental philosophical attacks on reliability. But things are somewhat more complicated.

3.4 Mentalism and Generalizability

Things are somewhat more complicated because philosophy is generally aimed at the articulation of a shared subject matter, one that can be conversed about with other philosophers, and is, at least in many cases, continuous with questions raised by non-philosophers (Jackson 1998, Alexander and Weinberg 2007).

For mentalists, the communal aspect of philosophy manifests itself in the commitment, not only to the illumination of an individual's conceptual competence, but of a more-or-less publicly shared concept, competence, or theory. Frank Jackson puts it this way: "Intuitions about how various cases, including various merely possible cases, are correctly described … are precisely what reveal our ordinary conceptions … or as it is often put nowadays, our folk theory of them" (1998, 31). For Jackson, the philosophical use of intuitions is continuous with the psychological study of common concepts. Crucially, once we allow that the philosophical object of inquiry is a *shared* or *average*, or *common* competence, mentalism's argument for the reliability of intuition starts to depend upon some principle of *Generalizability*. For example,

> *Generalizability of intuitions*: the intuitions of an individual philosopher (or a group of philosophers) are typical of a broader (or much broader, or very much broader . . .) class of people.

A principle like this would license the use of philosophical intuitions (obtained from an armchair, or a group of armchairs) to draw inferences about a broader community of thinkers. Another principle that might do would be:

> *Generalizability of competences*: the competences of an individual philosopher (or a group of philosophers) are typical of a broader (or much broader, or very much broader . . .) class of people.

This principle would allow us to generalize our best accounts of the competences underlying philosophers' intuitions to a broader community of thinkers.

But even if we grant the mentalist argument that there is a reliable connection between an individual's intuitions in good circumstances and that individual's conceptual competence, it does not ensure reliable knowledge of our object of inquiry since the various versions of Generalizability may be false.

Of course, it's always open to the mentalist to offer arguments to the effect that some version of Generalizability is true. For example, rationalists have suggested that some philosophical intuitions might be manifestations of an innate body of knowledge possessed by a person, and perhaps by every person. Noam Chomsky famously argued that humans possess an innate knowledge of principles of grammar for natural language (1986), a suggestion that has itself been the model for the further hypothesis that humans may possess an innate "moral grammar" that is causally implicated in the production of moral intuitions (Dwyer 1999, Harman 1999, Mikhail 2000, 2011). In a related vein, Jennifer Nagel (2012) has recently suggested that our knowledge of the epistemic domain is underwritten by our species-typical "theory of mind" capacities. [19] On this account, intuitions might be reliable evidence for the reconstruction of a widely shared competence, provided further arguments supporting the existence of shared, human-typical knowledge of the domain are sound. In any case, arguments for (some version of) Generalizability are substantial empirical claims in need of empirical support well beyond the sort that can be obtained from the armchair. For this reason, the perceived shortcomings of Jackson's (1998) defense of armchair intuitions as a proto-scientific way of discovering folk theories was and remains a major impetus for experimental philosophers worried about the extent to which Generalizability is true, especially given the narrow range of diversity within academic philosophy (Stich and Weinberg 2001; Buckwalter and Stich 2014). While some intuitive competences may be shared among all humans, we cannot generally discern which ones in advance of investigation (Henrich et al. 2010).[20] And the experimental philosophical work already reviewed gives preliminary evidence that diversity could turn out to be common, at least in some domains. Because the truth of (some suitable version of) Generalizability is a fact about the world "out there," versions of mentalism that are committed to it seem to enjoy no fundamental epistemic advantage over extramentalism. For both, evidence of diversity in intuitive responses undermines the reliability of these responses in indicating the truth about the object of study.

3.5 Interim Summary: Diversity and Reliability

We have suggested that experimental philosophical work be interpreted as challenging Reliable, and arguments from diversity provide an avenue for such a challenge that do not

[19] Cf. Stich 2013 and Nagel 2013 for ongoing debate of Nagel's view.

[20] This is true even if we allow Nagel's recent claim that for some novel judgments, confidence in the intuition is well correlated with the extent to which others in the subject pool share the intuition (2012). If (as I presume) these judgments emerge from one's experience of one's own community or language, the correlation will not hold across communal or linguistic borders. (Cf. Wright 2010, 2013.)

in any straightforward way fall into radical skepticism or require a more exacting standard than Reliable.

We have also suggested that diversity counts against reliability in somewhat different ways for extramentalists and mentalists. For extramentalists, disagreement among apparently epistemically parallel intuiters undermines our belief that intuitions that *p* are a reliable indicator of the truth of the intuition's content. Accumulating evidence of such disagreement can, more generally, undermine the presupposed disquotational account. For mentalists, diversity among apparently epistemically parallel intuiters undermines belief that philosophers' intuitions are reliable indicators of more widely shared intuitions, or of a more widely shared "folk" theory, conception, or concept in a domain.

4. Responding to Diversity

We noted in section 2.4 that there are at least three ways to respond to the challenge of diversity. One possibility—All Noise—is to take the failure of convergence of responses to experimental philosophical vignettes to indicate a general failure of judgments about cases to indicate any subject matter of concern. This could be the case, for example, if subjects' responses are determined by numerous subtle situational factors that swamp any effect of underlying conceptual competences. It could also be the case if the background conceptual competence is not structured, or simply doesn't exist, in the way philosophers imagine (Stich and Weinberg 2001).

However, many critics of experimental philosophy have pursued different, less radical approaches. The first, Noise and Signal response, challenges experimental philosophical work directly by claiming that the sorts of evidence that experimentalist critics advert to fails to detect the *signal* that is of philosophical interest, instead reflecting some sort of irrelevant *noise*, perhaps as the result of inadequate experimentation.[21] Alternatively, others have suggested that evidence of diversity in responses to vignettes may be evidence of Multiple Signals—in effect, that there are multiple different concepts or competencies at stake in producing the intuitions in play.

Elsewhere, I and others have noted that there are serious problems for deciding among these alternatives in particular cases, for deciding whether some behavioral disposition or response amounts to Noise or an expression of a conceptual competence (Machery 2008, Alexander et al. 2010a, 2010b). Because making such decisions seems to require individuating competences, something that there is no consensus on how to do (particularly if the competences are concepts), there remain substantial questions about how decisively discussion can be advanced. However, I note these worries only to put them aside here.

In this section, I discuss these Noise and Signal and Multiple Signals responses in general, and also in the particular case of intuitions about reference.

[21] Many experimental philosophers themselves have this same perspective, designing experiments that, they hope, capture what is of philosophical interest and not what is not. Some versions of the Naturalists' Challenge reflect this, challenging whether the armchair practitioner can separate the signal from the noise.

4.1 Noise not Signal

The most common way of critiquing experimental work is to suggest that it does not reveal the conceptual competence of underlying interest. And one common way of suggesting where it goes wrong is to suggest that an experiment elicits responses that reflect some variety of *pragmatic* or *speaker's meaning* rather than the literal or *semantic meaning* associated with the term or terms in the vignette (e.g. Adams and Steadman 2004, Deutsch 2009; Kauppinen 2007, Ludwig 2007, Sytsma and Livengood 2011).

One way of developing this response is to suggest that one or more experiments are badly designed (e.g. Sytsma and Livengood 2011), though some have gone further to suggest that psychological experimental methods are poorly suited to the enterprise of exploring concepts or word meanings (Kauppinen 2007, Ludwig 2007). The best response to the former critiques is for experimentalists to do more and better experiments, gradually controlling for the diverse sources of confusion that critics identify (as Justin Sytsma and Jonathan Livengood themselves undertake to do), and perhaps along the way this strategy will generate persuasive evidence that the latter critique is misguided.

As we noted, merely claiming that an experimental result is some sort of mistake or noise rather than a successful indication of the signal is not very satisfying in the absence of an account of the systematicity of the findings by experimental philosophers. Some critics have suggested, for example, that experimental philosophers' surprising findings of irrelevant determination stem from the fact that the experiments were carried out upon lay subjects instead of trained professional philosophers (Ludwig 2007, Devitt 2011, Pinillos et al. 2011)—a charge that has given rise to a lively debate concerning the possibility of philosophical expertise (Weinberg et al. 2010, Machery 2012). However this debate turns out, the critique here suggests strategies for demonstrating systematicity: the critic offers reasons why experimental results on everyday subjects might amount to mistakes not present in the population of professional philosophers.

In contrast, critics who suggest that cultural differences in responses to thought experiments are the result of error or pragmatic construal on the part of one cultural group or another need to offer an account of why such an error or construal should be found more systematically within one group than within another. Ideally, such an explanation is not offered post hoc, but is used to frame and test further hypotheses that could guide research to confirm or disconfirm it.

4.2 Diversity as Multiple Signals

A somewhat different style of response regards evidence of diversity, not as evidence that multiple, epistemically parallel actors are failing to respond to a common signal, but rather as evidence that there are multiple objects of knowledge—multiple signals—in play. Ernest Sosa (2007) has pushed this objection to experimental philosophers, noting that, "verbal disagreement *need* not reveal any substantive, real disagreement, if ambiguity and context might account for the verbal divergence" (p. 103).

Understanding diversity as merely a manifestation of verbal disagreement looks to have similar consequences for the disquotational/extramentalist accounts and the mentalist

accounts of philosophical method. Extramentalist accounts, recall, depend upon the reliability of one's Φ-capacity to indicate facts about the domain of the intuition. Crucially, if a diversity of responses to a survey reveals *conceptual* diversity (as Sosa's talk of "ambiguity and context" suggests), then this diversity does not undermine the evidential status of intuitions. Since apparent disagreements in uttered judgments will express attitudes towards distinct propositions, they need not entail any genuine disagreement. Thus, there need be no disagreement in intuitive judgments that could undermine the reliability of intuitions in providing evidence about the represented domain. For this reason, a Multiple Signals approach may be attractive to the extramentalist.

However, the cost to the extramentalist account of saving Reliable by endorsing Multiple Signals is *parochialism*. *Parochialist* accounts of philosophical intuition suggest that intuition aims at knowledge of a subject matter of no general interest, but only of interest to people who happen to have a particular, perhaps even idiosyncratic concept. The extramentalist who allows that to study knowledge is merely to study one property among numerous possible ones (say, knowledge*, and knowledge**) owes an explanation for why we ought to think that this property, the one that our own (perhaps cultural- or individual-bound) *knowledge* concept happens to give us reliable evidence for, is important (Stich 1990).

The mentalist pays a similar cost for embracing Multiple Signals, for it concedes that the psychological or mental competences that mentalists want to draw inferences about are, to some (perhaps significant) extent, not generalizable. Undertaking the study of the concept or conceptions of a single person or small group of people comes to look like a peculiarly limited form of ethnography, and one whose importance needs defending.

The resulting cost to both extramentalists and mentalists of saving Reliable by interpreting diversity as Multiple Signals looks to be some form of parochialism.

Relatedly, positing conceptual diversity, as Sosa does, seems to invite the possibility of widespread failures of communication.[22] Notice that interpreting divergent responses to thought experiments as evidence of divergent conceptual competences is a strategy that can be employed both across and *within* cultures.

Add to this that actual experimental philosophical surveys often exhibit a high degree of variation even within cultures, and we have preliminary evidence that conceptual diversity is widespread. But to the extent such diversity is widespread, a widening collapse of communication seems to loom. If *knowledge* concepts are as thinly sliced as cases of intuitive disagreement, then there will be many knowledge concepts, and much of our "knowledge" talk will simply be talking past one another.

4.3 The Case of Referential Intuitions

Consider again the case of intuitions about reference. Simple versions of descriptivist accounts of reference hold that the referent of a term is the thing that uniquely or best satisfies a description associated with a term by competent users (Lewis 1970; 1972). And simple versions of more recent, causal–historical accounts of reference (Kripke 1972/1980, Putnam 1975) hold that a term's referent is the thing that stands in an appropriate causal–historical

[22] Sosa recognizes this concern (2007, p. 103), but takes it to be of very limited application.

relationship to the term, allowing users' associated descriptions of the referent to be completely mistaken.

Following on work by Nisbett (2003) and Weinberg et al. (2001), Machery et al. (2004) found that Hong Kong subjects responded to vignettes concerning the reference of proper names with descriptivist answers while New Jersey subjects responded with causal–historical answers. In one study, they offered subjects vignettes modeled on a famous case from Saul Kripke. One example read as follows:

> Suppose that John has learned in college that Gödel is the man who proved an important mathematical theorem, called the incompleteness of arithmetic. John is quite good at mathematics and he can give an accurate statement of the incompleteness theorem, which he attributes to Gödel as the discoverer. But this is the only thing that he has heard about Gödel. Now suppose that Gödel was not the author of this theorem. A man called "Schmidt", whose body was found in Vienna under mysterious circumstances many years ago, actually did the work in question. His friend Gödel somehow got hold of the manuscript and claimed credit for the work, which was thereafter attributed to Gödel. Thus, he has been known as the man who proved the incompleteness of arithmetic. Most people who have heard the name "Gödel" are like John; the claim that Gödel discovered the incompleteness theorem is the only thing they have ever heard about Gödel. When John uses the name "Gödel", is he talking about:
>
> (A) the person who really discovered the incompleteness of arithmetic? or
> (B) the person who got hold of the manuscript and claimed credit for the work?

Because (A) picks out a person who satisfies John's description, but (B) picks out the person causally connected to John's use of "Gödel," it seems that A is more consistent with descriptivism and B with a causal–historical approach. Surprisingly (but consistent with their hypothesis), Machery et al. found that Hong Kong subjects were more likely to endorse descriptivist responses while New Jersey subjects' judgments were more consistent with a causal–historical understanding.

This work has now been subject to considerable additional discussion and criticism. Here I consider specifically how each of these two lines of critique (from 4.1 and 4.2) apply to this domain.

Almost from the beginning, Machery et al. (2004) has been criticized as failing to rule out the possibility of a *pragmatic* interpretation on which subjects offer a descriptivist response by way of picking out *what the speaker intends to refer to* rather than the *semantic reference*—what the word *literally* refers to (Kauppinen 2007, Ludwig 2007, Deutsch 2009, Sytsma and Livengood 2011; Ichikawa et al. 2012). It is certainly possible that this study indicates a difference between speakers' reference and semantic reference construals. But notice that this interpretation offers no clear explanation for why a cultural difference in such construals ought to have been found, and so it fails to explain the systematicity of such results. Subsequent work has attempted to further explore this hypothesis with some vindication for the finding of semantic diversity (Machery, Sytsma, and Deutsch 2015; Sytsma, Livengood, Sato, and Oguchi 2015). Other critics have worried that the results emerge from comparing native versus non-native speakers of English (Lam 2010), and again, while follow-up research has been limited, they offer some vindication for the original findings (Machery et al. 2010).

Given both its complexity and the empirical questions yet to be resolved, I will not adjudicate this debate further here. For our present purposes what is useful is that these

responses are illustrations of a Noise and Signal response. Critics are charging, in effect, that *semantic reference* is the signal, but that the findings of cultural diversity result in some way from some sort of irrelevant noise: perhaps a pragmatic construal in terms of speakers' reference, perhaps extra cognitive load from engaging in the task as a non-native speaker.

As I have argued, the second option, the Multiple Signals approach, risks both parochialism and communicative breakdown. Mallon et al. (2009) explore the possibility of interpreting diversity of referential intuitions in terms of multiple signals by considering the implications on *arguments from reference*. Arguments from reference are arguments that use the truth of a substantive account of the reference relation (for a class of terms) as a premise in an argument with a substantive philosophical conclusion. So, for example, eliminativists about propositional attitudes in the philosophy of mind (e.g. Churchland 1981), or about race in social philosophy (e.g. Appiah 1995; Zack 1993), have sometimes proceeded by specifying the content of a description that is purportedly associated with a term (e.g. "belief" or "race"), and then showing that nothing satisfies that description. They conclude that the term fails to refer, and that its referent does not exist. But this argument presupposes some simple version of descriptivism on which reference proceeds via satisfaction of the description associated with a term (Stich 1996). In contrast, others have noted that alternative, causal–historical accounts of reference allow successful reference even in the absence of successful description, and so eliminativism does not follow (Lycan 1988, Andreasen 1998).

Combining the Method of Cases with an (empirically possible and perhaps likely) diversity in intuitions about the reference of terms leads to the thought that a causal–historical theory of reference might be correct for some experimental subjects (most of the New Jersey students) and a descriptivist theory might be correct for others (most of the Hong Kong students). This is, in effect, to interpret diverse intuitions as Multiple Signals, each indicating a different, genuinely semantic notion revealing subjects' underlying concept of reference. But once we consider arguments from reference, this referential pluralism gives us the paradoxical result that the utterance "beliefs exist" might be true in New Jersey, but false in Hong Kong. What's gone wrong? One obvious solution is to relativize the reference of utterances to the theory of reference appropriate to the intuitions of the speaker. So, in this case, a New Jersey speaker who says that "beliefs exist" is referring to something different than a Hong Kong speaker who says that "beliefs do not exist." This invites parochialism and communicative breakdown.

In fact, as Mallon et al. argue, the situation may be much worse. If there is not only intercultural, but also *intra*cultural variation in intuitions about reference, then relativation opens up the possibility of parochialism and communicative breakdown among those with disparate intuitions within a culture. And this would be a striking consequence: utterances like "beliefs exist" or "races do not exist" could then have different truth values even for speakers within a culture whose intuitions about reference differ. Even more surprisingly, Mallon et al. argue, we may not even know the right interpretation for members of our own culture (or even ourselves). Consider that what the philosophical literature on the theory of reference shows is that the determination of a correct reference relation for a person or group is a *fine-grained* inquiry, one that requires consideration of a great many cases—the elicitation of a great many intuitions—to complete. But most of us have not undertaken such an inquiry, and indeed, it's not even clear that we yet know all the relevant questions to

ask, all the relevant thought experiments to pose. Combine the fine-grainedness of the project, the incompleteness of the project, and the diversity (both inter- and intra-culturally) in responses to thought experiments about reference. The upshot would be that we simply do not know which reference relation to use, for others or even for ourselves. So the "Multiple Signals" strategy of relativizing the interpretation of an utterance to the concept of reference that best fits the relevant intuitions of a speaker is one that suggests ubiquitous failures of communication. Mallon et al. argue that escaping this untenable consequence requires the abandonment of the use of intuitions as evidence for the truth of reference relations.[23,24]

4.4 Interim Summary

I have gone through the trouble to detail how the Noise and Signal strategy and the Multiple Signals strategy have played out in the debate over the experimental philosophy of reference in order to illustrate these responses, but also in order to set the stage for understanding how recent experimental results might change this dialectical landscape, both in the case of reference but perhaps also more broadly.

In section 5.1, I explore interpreting a diversity of responses as evidence of how multiple signals plays out, not only across cultures, and not only within cultures, but also within particular individuals. I present cases in which individuals construe extensionally equivalent vignettes in different ways, suggesting multiple conceptual competences in play within each individual. I then assess the impact on the dialectical landscape described thus far, using the case of reference as an illustration.

5. Intra-individual Noise and Ambiguity

Could diversity in responses to thought experiments be evidence of multiple conceptual competences even within individuals? What does it mean to describe an individual's responses as diverse in this way?

Intra-individual cases of variation in responses to thought experiments might be thought to be impossible to come by. Where an individual responds to equivalent materials in two different ways, it is plausible and tempting to interpret the individual as having made a mistake in at least one of the two cases (either All Noise, or Noise and Signal). Another possibility, however, is that the materials are construed ambiguously by the subject, in effect, activating more than one conceptual competence. In such cases, there need

[23] Ichikawa et al. 2012 offer a brief reply to these arguments. See Machery et al. 2013 for a response.

[24] Does this argument extend to other concepts of philosophical interest? To the extent that, say, a concept like *knowledge* is fine grained, then individuating the genuine concept from nearby alternatives (*knowledge*, knowledge***, etc.) will require consideration of a wide range of subtle and different cases. Again, plausibly, some of them may not even have been devised yet. And to the extent that intuitions about these cases vary, the Multiple Signals approach may mean a collapse of the publicity of concepts.

be no error. Instead, such cases could be instances of merely apparent intuitive conflict that can be settled by disambiguating the conceptual content of the intuitions (an intra-individual version of the Multiple Signals interpretation).

What would such a case look like?

5.1 Inference to Noise; Inference to Signals

We have said that framing effects sometimes are cases of extensionally equivalent descriptions. Consider Tversky and Kahneman's famous "Asian Disease Case" which offered subjects the following vignette:

> Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume the exact scientific estimate of the consequences of the programs are as follows.
>
> Subjects were divided into two conditions. In one condition, subjects were offered a choice between:
>
> Program A: "200 people will be saved"
>
> Program B: "There is a one-third probability that 600 people will be saved, and a two-thirds probability that no people will be saved"
>
> In the other condition, subjects were offered a choice between:
>
> Program C: "400 people will die"
>
> Program D: "there is a one-third probability that nobody will die, and a two-third probability that 600 people will die"

One thing to notice is that the average outcome of every choice is the same so that, if one were faced with the same problem repeatedly, each of the options would save the same number of people, on average. But more importantly, the two option sets describe equivalent possibilities in different ways. A is equivalent to C (if 200 people are saved then 400 die), and B to D (in each a 1/3 chance of saving everyone, and a 2/3 chance of saving no one). But what Tversky and Kahneman (1981) found was that subjects in the first condition overwhelmingly preferred program A to B (72% to 28%), but subjects in the second condition overwhelmingly preferred D to C (78% to 22%).

The Asian Disease Case reveals four things of use to us here.

First, intuitive judgments to hypothetical cases can differ between options that are extensionally equivalent.

Second, this difference is, plausibly, a result of different conceptual construals of the cases. On the standard understanding, the two ways of framing the situations described alter the subjects' construal of gains and losses, thereby changing their answers (because of greater loss aversion by subjects).

Third, plausibly, these mechanisms are partially *informationally encapsulated*. Information encapsulation occurs when a cognitive mechanism lacks access to or otherwise fails to make use of information that the organism as a whole possesses. Paradigmatically, visual illusions like Müller-Lyer do not dissipate, even when one knows that an illusion is present. The visual system continues to process the input in the same way, without making use of background information about properties of the input (Fodor 1984).

Similarly, framing effects can persist even when one knows of them. For example, as I switch back and forth between the two conditions, I find myself intuitively (in the "seeming" sense) preferring A in the first condition, and D in the second, even as I repeat the exercise, going back and forth. For this sort of reason, Tversky and Kahneman (1986) write that,

> On several occasions we presented both versions to the same respondents and discussed with them the inconsistent preferences evoked by the two frames. Many respondents expressed a wish to remain risk averse in the "lives saved" version and risk seeking in the "lives lost" version, although they also expressed a wish for their answers to be consistent. In the persistence of their appeal, framing effects resemble visual illusions more than computational errors. (S260)[25]

A consequence of this encapsulation is the existence of multiple points of intuitive pull—what we can call *intuitive disequilibrium*.[26] We all know what perceptual disequilibrium is like: we can experience it when we observe a Necker cube, and shift back and forth between perceiving one corner as foremost, and then another. Something like this happens in (at least some) framing effects as well, as when one feels the pull to switch back and forth between the A and D in the two conditions.

Fourth, as we suggested, cases of diverse responses to equivalent materials raise difficult questions about whether one response or another is in error.[27] In the vast literature on framing effects, many have suggested that studies like this show subjects exhibiting a kind of irrationality. Walter Sinnott-Armstrong even says that unreliability "follows from the very idea of framing effects" (2008, p. 52). Why should this be?

While most framing effects are exhibited between subjects, the different responses in different conditions by different subjects are assumed to reflect how individuals would respond in considering either case alone, and such counterfactual individuals seem irrational for making different choices in the two cases. For our present purposes, this interpretation of framing effects counts as either an All Noise or a Noise and Signal style interpretation since it views the conflict between conditions as evidence of underlying irrationality, undermining reliability in the domain (Sinnott-Armstrong 2008).

However, nothing necessitates this interpretation. In fact, we can interpret framing effect cases as evidence that Multiple Signals are operating. Once we do, we simply have different conceptual construals operating that may well rationally lead to different decisions (Van Roojen 1999; Frisch 1993).

Whether or not this classic framing effect can be successfully interpreted in terms of "Multiple Signals," it remains possible that some cases of intuitive disequilibrium might be so explained. These would be cases where extensionally equivalent cases are conceptually ambiguous, activating two conceptual competences. (This is simply to employ the same inference we have already applied, the inference from different responses to thought experiments to the successful operation of distinct underlying conceptual capacities.) In the next section, I develop an example of this alternate interpretation in more detail.

[25] Stanovich and West emphasize that more intelligent subjects presented with the cases together try to make their answers consistent, presumably recognizing that failure to do so would be irrational (1998).

[26] Gendler (2007) offers an interesting and suggestive discussion of intuitive disequilibrium.

[27] See Weigel (2012) for discussion of competing intuitions regarding free will.

5.2 Referential Ambiguity

We have noted recent research suggesting cross-cultural and even intra-cultural diversity with respect to intuitions about the reference of proper names. Recently, Shaun Nichols, Ángel Pinillos and I have been exploring the suggestion that diversity in referential intuitions might extend even into individual minds, in the form of "referential ambiguity." "Ambiguity" talk makes sense when we interpret multiple conceptual competences as operating in the same individual in response to equivalent stimuli. Our basic hypothesis is that descriptive and causal–historical accounts of reference might amount to characterizations of different conceptual competences—both present in individual minds—that may be activated in different circumstances.[28]

Using natural kind terms in cases like those from the history and philosophy of science, we hypothesized that if referential ambiguity exists, we might be able to manipulate such construals by individuals by manipulating non-representational features of the experimental materials. To test this hypothesis, we designed a series of experiments to "push around" intuitions between the two construals.

In one experiment, we offer subjects the following vignette:

> In the Middle Ages, animal researchers encountered a distinctive kind of bug. They noticed that these bugs seemed to be suspended in the air. Danish researchers called them "edderkops". It was thought that edderkops lived in the air and fed on the air. It was also thought that edderkops were worms that reproduced asexually.

We then added:

> (A1) However, it has turned out that there is nothing that has that collection of features.

Subjects were then asked the question: "What do you think about whether edderkops really exist?" and asked to indicate their response on a Likert scale from 1 (*edderkops* don't exist) to 6 (*edderkops* do exist). The mean response indicated agreement with the claim that edderkops do not exist (M = 1.94).

Because the description associated with the term turns out to be false, the descriptivist theory suggests that the term does not refer, and this mean response is consistent with the idea that descriptivism is (or is closer to) the folk theory of reference. In contrast, since the passage says that the scientists encountered a "distinctive kind" and called these bugs "edderkops," then it is plausible to think that the causal–historical theory predicts that the term *does refer* to the bugs, and that they *do exist*.

In second condition, however, we altered the question so as to existentially quantify over the kind:

> (A2) However, it has turned out that what the researchers called "edderkops" do not have that collection of features (and nothing else does either).

[28] Something very much like this view has been offered to account for natural kind terms in scientific discourse by Philip Kitcher (1993). Our recent work concerns folk intuitions about natural kind terms.

This second version of the final sentence restates what is already entailed by the vignette: namely, that there is a sort of thing that the researchers labeled. But responses to (A2)(M = 3.24) were significantly less descriptivist and more causal–historical than (A1). ($t(39) = 2.7$, $p<.01$).[29] One possibility is that subjects are simply making a mistake, that they answer differently to (A2) than (A1) because they quickly forget what the vignette said. However, other evidence supports an ambiguity view. For example, you can obtain a similar shift in existential judgments by simply priming subjects with an unrelated case of scientific error.[30]

In any case, this work is obviously just preliminary, and more empirical work needs to be done to confirm this hypothesis. But suppose that it is right. Suppose we do construe the responses to the vignettes we have described in terms of Multiple Signals, as talk of "referential ambiguity" suggests. What might follow from this?

5.3 Ameliorating Multiple Signals

Once we posit ambiguity, or conceptual diversity *within* the individual, the problems with Multiple Signals shift somewhat.

Consider first the problem of communicative breakdown. Even if divergent responses to thought experiments are expressions of different conceptual competences, no deep breakdown follows. So, for example, if Sid holds that "beliefs exist," and Nancy holds that "beliefs don't exist," they may be talking about different things (or expressing different concepts) with "belief." But on the ambiguity interpretation, Sid might still have the competence to make the judgment that Nancy is making, and Nancy that of Sid. Where multiple conceptual competences exist within individuals, additional conversation can resolve ambiguities, allowing for communication about those competences or the features of the world that they express or represent.[31] Such a process could reach as fine a grain as necessary.

The problems with parochialism are not so easily resolved.

Parochialism might initially seem less pressing. Worries about parochialism arise because diversity in intuitions suggests the possibility that an object of philosophical inquiry is not of very general interest, that what you are talking or thinking about and what I am talking or thinking about are entirely different. However, to the extent that diversity is a result of the differential operation of multiple conceptual competences already present within individual thinkers, it suggests that we each already have the conceptual repertoire appropriate to diverse domains. Sid might think that what Nancy talks about with "belief" is not important to talk about. But this won't fall out of a difference in their conceptual repertoires.

Still, the problem of parochialism persists. This is first because it is an open empirical question whether the sort of within-person diversity we have tried to illustrate in the case of reference is to be found in other domains where there is interpersonal diversity (though

[29] Nichols, Mallon, Pinillos (unpublished data).

[30] Nichols, Pinillos, Mallon (forthcoming) Experiment 1. For additional experimental evidence of the hypothesis, see Nichols, Pinillos, Mallon (forthcoming).

[31] The role of dialogue in determining the relevant concepts is emphasized by some critics of experimental philosophy (Kauppinen 2007, Ludwig 2007, Sosa 2007).

the possibility of genuine indeterminacy of folk competences is a standing possibility in philosophical discourse[32]). But it is more importantly because philosophical practice still needs defending.

Stephen Stich has pressed the problem of parochialism, asking: faced with a series of possible accounts of, say, belief or reference or knowledge, why should we think one is of more fundamental interest than another (Stich 1990, especially Chp. 5)? Sid does philosophical inquiry with or into his *belief*$_S$ concept, and Nancy with or into hers *belief*$_N$. But why should we take either inquiry to be worthwhile? This question is not answered by appealing to conceptual repertoires. Given all the concepts available to them, and especially the possibility that there are nearby concepts that do similar work in different ways—the question becomes only more pressing.

An old answer to Stich's question was: one relation, property, or concept is of more interest than another because that is the one specified by our common sense. Evidence for intuitive diversity was initially offered as a challenge precisely to this sort of view (Weinberg et al. 2001, Nichols et al. 2003). If we allow that finding conceptual diversity within individual minds undercuts the charge of parochialism (since many intuitions will equally have a claim to being expressions of folk conceptual competences), then it equally undercuts the original response to Stich since common sense would no longer specify a unique relation, property, or concept. The indeterminacy that exists among the folk conceptions of different people is simply moved inside the head. We remain left having to ask: why this one rather than that one?

6. Going Forward with Experimental Philosophy

Experimentalist extensions of the Naturalists' Challenge have, and continue, to find ways to challenge traditional philosophical methods through the use of experimental technique. The result is a complex landscape of challenges and replies, and here I have simply sketched one path through a complex dialectical thicket.

I argued that the challenge of experimental philosophy seems in the first place to be a challenge to the reliability of intuition, but defenders of arguments from irrelevant determination need to do more if they are to challenge Reliable. In contrast, arguments from diversity seem capable of posing a challenge to Reliable. Systematic intuitive disagreement among apparently epistemically parallel actors poses a prima facie challenge to the reliability of intuitions, whether they are concerned with indicating the features of the world they represent or with indicating the putatively shared conceptual competences that that they express.

Sosa argued, correctly, that intuitive disagreement (understood as divergent responses to experimental surveys) may not indicate genuine disagreement, but rather might indicate verbal disagreement. Interpreted in this way, diversity need not pose a prima facie challenge

[32] See, e.g. Sider 2001 on personal identity, or Ned Hall (2004) on causation.

to Reliable. But interpreting diversity in this way risks making philosophical inquiry parochial and suggests also the possibility of widespread communicative breakdown.

In the last section, I also explored the application of Sosa's reasoning to intra-individual diversity exhibited in responses to framing effects, and vignettes modeled on cases from the history and philosophy of science. In some such cases, it seems plausible to interpret responses to extensionally equivalent vignettes as evidence of diverse conceptual construals of the vignette. Where it is appropriate to do so, the problem of conceptual breakdown seems less pressing, though the questions that gave rise to parochialism remain as pressing as ever.

Looked at as a whole, it seems there's a plausible tradeoff between the scope of intuitions and their reliability. We saw this initially in the mentalist thought that intuitions might be better evidence of features of the human mind than they are of the facts of the world their contents describe. And we see it again in the trade-off that occurs when we preserve reliability by interpreting divergent intuitions as instances of multiple philosophical capacities, whether across cultures or within individual minds.

Is there more experimental progress to be made? The answer is certainly "yes." Further articulation of the psychological sources of intuitions and the factors that alter them will almost certainly better illuminate the extent to which various kinds of intuition ought to be regarded as errors or could plausibly be regarded as appropriate expressions of underlying competences. Further cross-cultural work will help make clearer the extent to which diverse responses to philosophical thought experiments poses a challenge to the Method of Cases. And further work manipulating the intuitions of particular individuals, should reveal the extent to which each individual's conceptual repertoire might have multiple conceptual competences that can be brought to bear on individual cases, perhaps illuminating longstanding philosophical disagreements.

Here I have focused primarily on experimental work growing out of the Naturalists' Challenge to the role of armchair intuitions The debates surrounding the Naturalists' Challenge exhibit the way that new empirical methods and results can drive philosophical inquiry in productive ways. But, as I noted at the outset, experimental philosophy is much larger than this dialectic, and quickly becoming larger still, blurring the disciplinary boundaries between psychology and neuroscience with consequences that have yet to reveal themselves. These developments will surely raise new questions and drive new debates in the years to come. Ultimately, what many experimental philosophers hope is that the use of experimental techniques is not a fad, but instead an opportunity to add novel methods to the broad disciplinary toolbox of philosophy in the way that, for example, formal logic, game theory, history, and Bayesian probability have already been added. If this does happen, then these are still very early days for experimental philosophy indeed.

References

Adams, F. and A. Steadman (2004). "Intentional Action in Ordinary Language: Core Concept or Pragmatic Understanding." *Analysis* 64: 173–81.

Alexander, J. (2012). Experimental Philosophy: An Introduction, Polity Press.

Alexander, J. and J. Weinberg (2007). "Analytic Epistemology and Experimental Philosophy." *Philosophy Compass* 2(1): 56–80.

Alexander, J., R. Mallon, et al. (2010a). "Accentuate the Negative." *Review of Philosophy and Psychology* 1(2): 297–314.

Alexander, J., R. Mallon, et al. (2010b). "'Competence: What's In? What's Out? Who Knows?' Commentary on Joshua Knobe's 'Person as Scientist, Person as Moralist'."*Behavioral and Brain Sciences* 33: 329–30.

Andreasen, R. O. (1998). "A New Perspective on the Race Debate." *British Journal of the Philosophy of Science* 49: 199–225.

Appiah, K. A. (1995). The Uncompleted Argument: Du Bois and the Illusion of Race. In *Overcoming Racism and Sexism*. L. A. Bell and D. Blumenfeld. Lanham, MD, Rowman and Littlefield: 59–77.

Bealer, G. (1998). Intuition and the Autonomy of Philosophy. In *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*. M. R. DePaul and W. Ramsey. Lanham, MD, Rowman and Littlefield: 201–39.

Block, N. (1978). Troubles with Functionalism. In *Perception and Cognition: Issues in the Foundations of Psychology: Minnesota Studies in the Philosophy of Science*, Vol. 9, C. W. Savage. Minneapolis, University of Minnesota Press: 261–325.

Buckwalter, W. and S. Stich (2014). Gender and Philosophical Intuition. In *Experimental Philosophy,* Vol. 2, J. Knobe and S. Nichols. Oxford University Press: 307–46.

Cappelen, H. (2012). *Philosophy Without Intuitions*. Oxford, Oxford University Press.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. New York, Oxford University Press.

Chomsky, N. (1986). *Knowledge of Language*. New York, Praeger.

Churchland, P. (1981). "Eliminative Materialism and the Propositional Attitudes." *Journal of Philosophy* LXXVII(2): 67–90.

Cokely, E. T. and A. Feltz (2009). "Adaptive Variation in Judgment and Philosophical Intuition." *Consciousness and Cognition* 18(1): 356–58.

Cullen, S. (2010). "Survey-Driven Romanticism." *Review of Philosophy and Psychology* 1(2): 275–96.

Cummins, R. (1998). Reflection on Reflective Equilibrium. In *Rethinking Intuition*. M. R. DePaul and W. Ramsey. Lanham, MD, Rowman and Littlefield: 113–28.

Davidson, D. (1987). "Knowing One's Own Mind." *Proceedings and Addresses of the American Philosophical Association* 60: 441–58.

Deutsch, M. (2009). "Experimental Philosophy and the Theory of Reference." *Mind and Language* 24(4): 445–66.

Deutsch, M. (2010). "Intuitions, Counter-Examples, and Experimental Philosophy." *Review of Philosophy and Psychology* 3(3): 447–60.

Devitt, M. (2006). *Ignorance of Language*. Oxford, Clarendon.

Devitt, M. (2011). "Experimental Semantics." *Philosophy and Phenomenological Research* LXXXII(2): 418–35.

Dorr, C. (2010). "Review of *Every Thing Must Go: Metaphysics Naturalized* by James Ladyman and Don Ross, with David Spurrett and John Collier, Oxford OUP, 2007." *Notre Dame Philosophical Reviews*. See <http://ndpr.nd.edu/news/24377/?id=19947>. Accessed September 23, 2015.

Dwyer, S. (1999). Moral Competence. In *Philosophy and Linguistics*. K. Murasugi and R. Stainton. Boulder, CO, Westview Press: 169–90.

Feltz, A. and E. Cokely (2012). "The Philosophical Personality Argument." *Philosophical Studies* 161(2): 227–46.

Feltz, A. and E. T. Cokely (2008). "The Fragmented Folk: More Evidence of Stable Individual Differences in Oral Judgments and Folk Intuitions." *Proceedings of the 30th Annual Conference of the Cognitive Science Society*. B. C. Love, K. McRae, and V. M. Sloutsky. Cognitive Science Society: 1771–6.

Fodor, J. (1984). "Observation Reconsidered." *Philosophy of Science* 51: 23–43.

Frisch, D. (1993). "Reasons for Framing Effects." *Organizational Behavior and Human Decision Processes* 54: 399–429.

Gendler, T. S. (2007). Philosophical Thought Experiments, Intuitions, And Cognitive Equilibrium. In *Philosophy and the Empirical*, Vol. 31. P. A. French and H. K. Wettstein. Oxford, Blackwell: 68–89.

Gettier, E. (1963). "Is Justified True Belief Knowledge?" *Analysis* 23(6): 121–3.

Glasgow, J. (2008). "On the Methodology of the Race Debate: Conceptual Analysis and Racial Discourse." *Philosophy and Phenomenological Research* 76(2): 333–58.

Goldman, A. (1976). "Discrimination and Perceptual Knowledge." *Journal of Philosophy* 73: 771–91.

Goldman, A. (2010). "Philosophical Naturalism and Intuitional Methodology (Romnell Lecture)." *Proceedings and Addresses of the American Philosophical Association* 84(2): 115–50.

Goldman, A. I. and J. Pust (1998). Philosophical Theory and Intuitional Evidence. In *Rethinking Intuition: The Psychology of Intuition and its Role in Philosophical Inquiry*. M. Depaul and W. Ramsey, Rowman and Littlefield: 179–97.

Greene, J. and J. Haidt (2002). "How (and Where) Does Moral Judgment Work?." *Trends in Cognitive Sciences* 6(12): 517–23.

Greene, J. D. (2003). "From Neural 'Is' To Moral 'Ought': What Are The Moral Implications Of Neuroscientific Moral Psychology?" *Nature Reviews Neuroscience* 4: 847–50.

Greene, J. D. (2007). The Secret Joke of Kant's Soul. In *Moral Psychology*, Vol. 3. W. Sinnott-Armstrong. Cambridge, MA, MIT Press: 35–117.

Greene, J. D., R. B. Sommerville, et al. (2001). "An fMRI Investigation of Emotional Engagement in Moral Judgment." *Science* 293: 2105–8.

Haidt, J. (2001). "The Emotional Dog and its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review* 108: 814–34.

Hall, N. (2004). Two Concepts Of Causation. In *Causation And Counterfactuals*. J. Collins, N. Hall, and L. Paul. Cambridge, MA, The MIT Press: 225–76.

Harman, G. (1999). Moral Philosophy and Linguistics. *Proceedings of the 20th World Congress of Philosophy, vol. I: Ethics*. K. Brinkmann. Bowling Green, Ohio, Philosophy Documentation Center: 107–15.

Henrich, J., S. J. Heine, et al. (2010). "The Weirdest People in the World." *Behavioral and Brain Sciences* 33: 61–135.

Ichikawa, J. (2011). "Experimentalist Pressure Against Traditional Methodology." *Philosophical Psychology* 25(5): 743–65.

Ichikawa, J., I. Maitra, et al. (2012). "In Defense of a Kripkean Dogma." *Philosophy and Phenomenological Research* 85(1): 55–68.

Jackson, F. (1982). "Epiphenomenal Qualia." *Philosophical Quarterly* 32: 127–36.

Jackson, F. (1998). *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Oxford, Oxford University Press.

Joyce, R. (2006). *The Evolution of Morality. Life and Mind*. Cambridge, MA; London, MIT Press.

Kauppinen, A. (2007). "The Rise and Fall of Experimental Philosophy." *Philosophical Explorations* 10(2): 95–118.
Kitcher, P. (1993). *The Advancement of Science: Science Without Legend, Objectivity Without Illusions*. New York, Oxford University Press.
Knobe, J. and S. Nichols (2008). Experimental Philosophy: A Manifesto. In *Experimental Philosophy*. J. Knobe and S. Nichols. New York, Oxford University Press: 3–16.
Kornblith, H. (2007). "Naturalism and Intuitions." *Grazer Philosophische Studien* 74: 27–49.
Kripke, S. A. (1980). *Naming and Necessity*. Cambridge, Massachusetts, Harvard University Press.
Lam, B. (2010). "Are Cantonese Speakers Really Descriptivists? Revisiting Cross-Cultural Semantics." *Cognition* 115: 320–9.
Lehrer, K. (1990). *Theory of Knowledge*. Boulder, Westview Press.
Lewis, D. (1970). "How to Define Theoretical Terms." *Journal of Philosophy* 67: 426–46.
Lewis, D. (1972). "Psychophysical and Theoretical Identifications." *Australasian Journal of Philosophy* 50: 249–58.
Ludwig, K. (2007). "The Epistemology of Thought Experiments: First Person versus Third Person Approaches." *Midwest Studies in Philosophy* XXXI: 128–59.
Ludwig, K. (2010). "Intuitions and Relativity." *Philosophical Psychology* 23(4): 427–45.
Lycan, W. (1988). *Judgement and Justification*. Cambridge, Cambridge University Press.
Machery, E. (2008). "The Folk Concept of Intentional Action: Philosophical and Experimental Issues." *Mind and Language* 23(2): 165–89.
Machery, E. (2012). "Expertise and Intuitions about Reference." *Theoria* 27(1): 37–54.
Machery, E., R. Mallon, et al. (2004). "Semantics, Cross-Cultural Style." *Cognition* 92: B1–B12.
Machery, E., M. Deutsch, et al. (2010). "Semantic Intuitions: Reply to Lam." *Cognition* 117(3): 361–6.
Machery, E., R. Mallon, S. Nichols, and S. P. Stich (2013). "If Folk Intuitions Vary, Then What?" *Philosophy and Phenomenological Research* 86(3): 618–35.
Machery, E., J. Sytsma, and M. Deutsch (2015). Speaker's Reference and Cross-Cultural Semantics. In *On Reference*. A. Bianchi. New York, Oxford University Press: 62–76.
McMahan, J. (2000). Moral Intuition. In *Blackwell Guide to Ethical Theory*. H. LaFollette. Oxford, Blackwell.
Mallon, R., E. Machery, et al. (2009). "Against Arguments from Reference." *Philosophy and Phenomenological Research* 79(2): 332–56.
Mallon, R. and J. Doris (2013). The Science of Ethics. In *Blackwell Companion to Ethics*. H. LaFollettte. Oxford, Blackwell: 169–96.
Mason, K. (2011). "Moral Psychology And Moral Intuition: A Pox On All Your Houses." *Australasian Journal of Philosophy* 89(3): 441–58.
Mikhail, J. (2000). Rawls' Linguistic Analogy: A Study of the "Generative Grammar" Model of Moral Theory Described by John Rawls in *A Theory of Justice*. Ph.D. Dissertation in Philosophy. Ithaca, Cornell University.
Mikhail, J. M. (2011). *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*. New York, Cambridge University Press.
Murray, D. and E. Nahmias (2014). "Explaining Away Incompatibilist Intuitions." *Philosophy and Phenomenological Research* 88(2): 434–67.
Nadelhoffer, T. (2006a). "Bad Acts, Blameworthy Agents, and Intentional Actions: Some Problems for Jury Impartiality." *Philosophical Explorations* 9(2): 203–20.
Nadelhoffer, T. (2006b). "On Trying to Save the Simple View." *Mind & Language* 21(5): 565–86.

Nadelhoffer, T. and E. Nahmias (2007). "The Past and Future of Experimental Philosophy." *Philosophical Explorations* 10(2): 123–49.
Nagel, J. (2012). "Intuitions and Experiments: A Defense of the Case Method in Epistemology." *Philosophy and Phenomenological Research* 85(3): 495–527.
Nagel, J. (2013). "Defending the Evidential Value of Epistemic Intuitions: A Reply to Stich." *Philosophy and Phenomenological Research* 86(1): 179–99.
Nagel, J., V. San Juan, and R. Mar. (2013). "Lay Denial of Knowledge for Justified True Beliefs." *Cognition* 129: 652–61.
Nahmias, E. A. and M. Thompson (2014). A Naturalistic Vision of Free Will. In *Current Controversies in Experimental Philosophy*. E. Machery and E. O'Neill. Chicago, Routledge: 86–103.
Nichols, S., S. P. Stich, et al. (2003). Metaskepticism: Meditations in Ethno-Epistemology. In *The Skeptics*. S. Luper. Burlington, VT, Ashgate: 227–48.
Nichols, S. and J. Knobe (2007). "Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions." *Nous* 41: 663–85
Nichols, S. and J. Ulatowski (2007). "Intuitions and Individual Differences: The Knobe Effect Revisited." *Mind and Language* 22(4): 346–65.
Nichols, S., À. Pinillos, and R. Mallon (forthcoming). "Referential Ambiguity." *Mind*.
Nisbett, R. E. (2003). *The Geography of Thought: How Asians and Westerners Think Differently—and Why*. London; Yarmouth, Maine, Nicholas Brealey Publishing.
Petrinovich, L., P. O'Neill, et al. (1993). "An Empirical Study of Moral Intuitions: Toward An Evolutionary Ethics." *Journal of Personality and Social Psychology* 64(3): 467–78.
Pinillos, Á., N. Smith, et al. (2011). "Philosophy's New Challenge: Experiments and Intentional Action." *Mind and Language* 26(1): 115–39.
Prinz, J. (2007). *The Emotional Construction of Morals*. New York, Oxford University Press.
Putnam, H. (1975). "The Meaning of 'Meaning'." *Minnesota Studies in the Philosophy of Science* 7: 131–93.
Rawls, J. (1971). *A Theory of Justice*. Cambridge, Mass., Harvard University Press.
Roskies, A. L. and S. Nichols (2008). "Bringing Moral Responsibility Down to Earth." *Journal of Philosophy* 105: 371–88.
Schnall, S., J. Haidt, et al. (2008). "Disgust as Embodied Moral Judgment." *Personality and Social Psychology Bulletin* 34(8): 1096–109.
Schwitzgebel, E. and F. Cushman (2012). "Expertise in Moral Reasoning? Order Effects on Moral Judgment in Professional Philosophers and Non-Philosophers." *Mind and Language* 27(2): 135–53.
Searle, J. (1980). "Minds, Brains and Programs." *Behavioral and Brain Sciences* 3(3): 417–57.
Seyedsayamdost, H. (2015). "On Normativity And Epistemic Intuitions: Failure Of Replication." *Episteme* 12(1): 95–116.
Sider, T. (2001). "Criteria of Personal Identity and the Limits of Conceptual Analysis." *Philosophical Perspectives* 15(s15): 189–209.
Singer, P. (2005). "Ethics and Intuitions." *Journal of Ethics* 9(3–4): 331–52.
Sinnott-Armstrong, W. (2008). Framing Moral Intuitions. In *Moral Psychology*, Vol. 2. W. Sinnott-Armstrong. Cambridge, MA, MIT Press: 47–76.
Sosa, E. (2007). "Experimental Philosophy and Philosophical Intuition." *Philosophical Studies* 132(1): 99–107.
Stanovich, K. E. and R. F. West (1998). "Individual Differences in Framing and Conjunction Effects." *Thinking and Reasoning* 4(4): 289–317.

Stich, S. (1990). *The Fragmentation of Reason: A Preface to a Pragmatic Theory of Cognitive Evaluation*. Cambridge, Massachusetts, MIT Press, A Bradford Book.
Stich, S. P. (1996). *Deconstructing the Mind*. New York, Oxford University Press.
Stich, S. (2013). "Do Different Groups Have Different Epistemic Intuitions? A Reply to Jennifer Nagel." *Philosophy and Phenomenological Research* 87(1): 151–78.
Stich, S. and J. Weinberg (2001). "Jackson's Empirical Assumptions." *Philosophy and Phenomenological Research* 62(3): 637–43.
Stotz, K., P. E. Griffiths and R. D. Knight (2004). "How Biologists Conceptualize Genes: An Empirical Study." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 35(4): 647–73.
Swain, S., J. Alexander, et al. (2008). "The Instability of Philosophical Intuitions: Running Hot and Cold On True Temp." *Philosophy and Phenomenological Research* LXXVI(1): 138–55.
Sytsma, J. and J. Livengood (2011). "A New Perspective Concerning Experiments in Semantic Intuitions." *Australasian Journal of Philosophy* 89(2): 315–32.
Sytsma, J., J. Livengood, R. Sato, and M. Oguchi (2015). "Reference in the Land of the Rising Sun: A Cross-cultural Study on the Reference of Proper Names." *Review of Philosophy and Psychology* 6(2): 213–30.
Thomson, J. J. (1985). "The Trolley Problem." *Yale Law Journal* 94(6): 1395–415.
Turri, J. (2013). "A Conspicuous Art: Putting Gettier to the Test." *Philosophers' Imprint* 13(10): 1–16.
Tversky, A. and D. Kahneman (1981). "The Framing of Decisions and the Psychology of Choice." *Science* 211(4481): 453–58.
Tversky, A. and D. Kahneman (1986). "Rational Choice and the Framing of Decisions." *Journal of Business* 59: 251–78.
van Roojen, M. (1999). "Reflective Moral Equilibrium and Psychological Theory." *Ethics* 109(4): 846–57.
Weigel, C. (2011). "Distance, Anger, Freedom: An Account of the Role of Abstraction in Compatibilist and Incompatibilist Intuitions." *Philosophical Psychology* 24(6): 803–23.
Weigel, C. (2012). "Experimental Evidence for Free Will Revisionism." *Philosophical Explorations* 16(1): 31–43.
Weinberg, J. (2007). "How to Challenge Intuitions Empirically Without Risking Skepticism." *Midwest Studies in Philosophy* XXXI: 318–43.
Weinberg, J. M. (2009). "On Doing Better, Experimental-Style. Review of Timothy Williamson, *The Philosophy of Philosophy*." *Philosophical Studies* 145(3): 455–64.
Weinberg, J., S. Nichols, and S. Stich. (2001). "Normativity and Epistemic Intuitions." *Philosophical Topics* 29(1&2): 429–59.
Weinberg, J. M., C. Gonnerman, et al. (2010). "Are Philosophers Expert Intuiters?" *Philosophical Psychology* 23(3): 331–55.
Weinberg, J., J. Alexander, et al. (2012). "Restrictionism and Reflection: Challenge Deflected, or Simply Redirected?." *The Monist* 95: 200–22.
Williamson, T. (2007). *The Philosophy of Philosophy*. Malden, MA, Blackwell Pub.
Williamson, T. (2013). "Review of J. Alexander, *Experimental Philosophy: An Introduction*." *Philosophy* 88(3): 467–74.

Wright, J. (2010). "On Intuitional Stability: The Clear, the Strong, and the Paradigmatic." *Cognition* 115(3): 419–503.
Wright, J. (2013). "Tracking Instability in Our Philosophical Judgments: Is it Intuitive?" *Philosophical Psychology* 26(4): 485–501.
Zack, N. (1993). *Race and Mixed Race*. Philadelphia, Temple University Press.

CHAPTER 23

TRANSCENDENTAL ARGUMENTS

DERK PEREBOOM

1. Introduction

Among Immanuel Kant's (1724–1804) most influential contributions to philosophy is his development of the transcendental argument. In Kant's conception, an argument of this kind begins with a compelling first premise about our thought, experience, knowledge, or practice, and then reasons to a conclusion that is a substantive and unobvious presupposition and necessary condition of the truth of this premise, or as he sometimes puts it, of the possibility of this premise's being true. Transcendental arguments are typically directed against skepticism of some kind. For example, Kant's Transcendental Deduction targets Humean skepticism about the applicability of a priori metaphysical concepts, and his Refutation of Idealism takes aim at skepticism about an external world. A focus on anti-skeptical objectives suggests that this method addresses only a fairly narrow range of philosophical topics. However, many issues in philosophy can be represented as confrontations between skeptical and anti-skeptical points of view. For example, the utilitarian can be represented as a skeptic about various non-consequentialist moral considerations, and the incompatibilist about free will and determinism as a skeptic about free will given determinism. Yet at the same time it is not essential to transcendental arguments that they have anti-skeptical intent.

2. The Nature of Transcendental Arguments

An important issue for transcendental arguments concerns the epistemic qualifications of the initial premise. It is sometimes specified that the first premise of a successful transcendental argument must be one of which we—and this includes a targeted skeptic—are or can be certain. But whether this standard must be met depends on the skepticism the argument

targets. Janet Broughton (2002) interprets Descartes's *cogito ergo sum* as a transcendental argument against the dream and evil demon skepticism introduced in the First Meditation. Given that the standard Descartes sets for the acceptability of a claim is indubitability, the initial premise must also be indubitable. But not every transcendental argument is advanced in such an epistemically rarified context. For instance, Justin Coates (forthcoming) provides a compelling interpretation of P. F. Strawson's 'Freedom and Resentment' (1962) as a transcendental argument against skepticism about moral responsibility. This skeptic is concerned that moral responsibility is incompatible with determinism and that determinism might well be true. Skepticism of this sort does not appeal to the claim that only indubitable propositions are acceptable, and consequently the first premise need not meet this standard.

An alternative and plausible standard is contextual—the first premise must be one the skeptic at issue will accept. It would be valuable in addition if this premise had a particular sort of resilience for the skeptic, so that once she understands that a necessary condition of the truth of that premise is a claim she doubts or denies, she won't be readily disposed to respond by denying the premise. This point might be made more precise when stated in terms of credences, that is, degrees of belief in a proposition or beliefs in probabilities that a proposition is true. If the transcendental argument is to succeed against the skeptic, the skeptic's credence in the first premise conditional on its necessitating the falsity of the skeptical claim must be not significantly lower than the skeptic's initial credence in the skeptical claim. True, no actual skeptic is likely to be convinced to reject his skepticism by a transcendental argument, human nature being what it is, and he will surely find fault with either the first premise or the reasoning. For this reason, 'the skeptic' in this characterization must be relevantly idealized.

Some transcendental arguments gain strength by appealing to a first premise that is a supposition the skeptic must accept either because it is a premise of the skeptic's argument for her skeptical conclusion, or because it is transparently entailed or presupposed by such a premise. If such a transcendental argument is successful, the ground of the falsity of the skeptic's claim would turn out to be a premise the skeptic cannot reject while retaining her skepticism. The resilience of such a premise would be especially strong. Any reduction of the skeptic's credence in the premise upon understanding that it necessitates the falsity of the skeptical claim would result in a corresponding weakening of the skeptic's argument, and in particular a reduction in her rational credence in the skeptical conclusion.

The crucial steps in the reasoning featured in transcendental arguments are claims to the effect that a subconclusion or conclusion is a presupposition and a necessary condition of a premise—that is, that the premise presupposes and necessitates the subconclusion or conclusion. On one proposal, these steps must display logically necessary conditions in particular, and weaker connections are insufficient. But here too it's plausible that the requisite strength of the necessary connection varies with the type of skeptic targeted by the transcendental argument. The necessity might be logical, metaphysical, nomological, or explanatory. If the skeptic doubts metaphysical necessitation but not logical necessitation, the necessary conditions appealed to in the argument must be logical. If the skeptic does not doubt either logical or metaphysical necessitation, then both logical and metaphysical necessary conditions are fair game. But the moral responsibility skeptic, for instance, takes no issue with nomological necessitation, and thus a transcendental argument that takes

aim at this position is free to employ such a weaker condition, and arguably Strawson's (1962) version does (Coates, forthcoming). Furthermore, in many philosophical contexts, the relevant sort of skeptic takes no issue with the notion of only possible explanation or best explanation. In such a context, the steps of a transcendental argument need show only that the subconclusion or conclusion is a necessary condition for the premise in the sense that it is the only possible explanation for it, or in the still weaker sense that it is the best explanation for it. As we shall see, the key steps of Kant's Transcendental Deduction invoke such explanatory conditions. This is not a defect in the argument, for the reason that Humean skepticism about the applicability of metaphysical concepts is not also skeptical of explanatory necessary conditions.

Perhaps the best-known contemporary transcendental arguments are *world-directed* in the sense that they aim to secure an anti-skeptical claim about mind-independent reality (Peacocke 1989; Cassam 1999; Stern 2012). But transcendental arguments need not be world-directed in this sense. They might, for example, be ethical in import without aiming to establish a moral realist conclusion, by contrast with one that is, say, constructivist instead. For example, Strawson's transcendental argument against skepticism about moral responsibility leaves a constructivist account of moral responsibility open. Christine Korsgaard's (1996) transcendental argument for the conclusion that we must value ourselves as rational agents from the premise that we make rational choices also does not commit to moral realism. As we shall see, the world-directed transcendental arguments face an important objection, advanced by Barry Stroud (1968), according to which the *existence* of the external feature will not be a necessary condition of the aspect of experience or knowledge invoked by the first premise, since a belief about or representation of the external feature would also suffice. A world-directed transcendental argument vulnerable to this objection would fall short of its anti-skeptical ambitions.

Let us now inspect a number of specific transcendental arguments, two from Kant and several contemporary examples. We will begin with a substantial discussion of Kant's Transcendental Deduction. It's still the paradigmatic transcendental argument, and due to its ambitiousness and promise, it has been the main inspiration for the ensuing tradition. We then turn to Kant's Refutation of Idealism, because it inspires the widespread strategy of using transcendental arguments to undermine external-world skepticism. Subsequently we will discuss a number of contemporary arguments, focusing on their problems and prospects.

3. Kant's Transcendental Arguments

Kant's most famous transcendental arguments are found in the *Critique of Pure Reason* (1781, 1787/1987): the Transcendental Deduction of the Categories, the Second Analogy, and the Refutation of Idealism. There are many others, in, for example, the *Critique of Practical Reason,* the *Critique of the Power of Judgment*, and in the *Opus Posthumum* (Forster 1989). Here I single out the two that are most celebrated: the Transcendental Deduction and the Refutation of Idealism. Discussion of the Transcendental Deduction, the most influential of all, and the part of Kant's theoretical philosophy that he believed to be his greatest

achievement, illustrates the structure of a transcendental argument, and in particular the epistemic requirements for the first premise and for the necessary conditions such an argument involves. Consideration of the Refutation of Idealism highlights in addition the type of objection Stroud raises against world-directed transcendental arguments.

3.1 The Transcendental Deduction

In the Transcendental Deduction (1781/1787/1987: A84–130, B116–169) Kant aims to demonstrate against an empiricist that certain priori concepts legitimately apply to objects featured in our experience. A deduction in this context is an argument intended to justify the use of a concept, one that shows that the concept legitimately applies to real things. For Kant a concept is a priori just in case its source is in the mind of the subject and not in sensory experience (A80/B106; Strawson 1966: 86). The particular a priori concepts whose applicability to objects of experience Kant aims to vindicate are given in his Table of Categories (A80/B106); they are *unity, plurality*, and *totality* (the Categories of Quantity); *reality, negation*, and *limitation* (the Categories of Quality); *inherence and subsistence, causality, and dependence*, and *community* (the Categories of Relation), and *possibility–impossibility, existence–non-existence, necessity–contingency* (the Categories of Modality).

David Hume denies that a deduction can be provided for a number of metaphysical concepts—*ideas*, in his terminology—including those of personal identity, of identity over time more generally, of the self as a subject distinct from its perceptions, and of causal power or force (1739, 1748). In his view, a concept can be validated only by finding a sensory experience, that is, an impression, in particular one that is the 'original' of that idea, which must resemble the idea. However, because any attempt to locate, for example, an impression corresponding to the idea of causal power turns out to be unsuccessful, he concludes that this idea does not apply in our experience (1748: §7). In Kant's terminology, Hume is testing to see whether there is an *empirical* deduction of the concept of causal power (A85/B117), and from the failure of the attempt to produce one, he concludes that this concept lacks *objective validity*, that is, it does not apply to the objects of our experience.

Hume's view about the impossibility of a deduction of a priori metaphysical concepts is Kant's target in the Transcendental Deduction. Kant agrees with Hume, however, that no empirical deduction is forthcoming for such concepts. Instead, he aims to produce a different sort of justification for them, one that is transcendental rather than empirical. A transcendental deduction begins with a premise about any possible human experience, a premise to which the participants in the debate will at least initially agree, and then contends that a presupposition of and necessary condition for the truth of that premise (or for the possible truth of that premise) is the applicability of the a priori concepts at issue, the categories. Kant's Transcendental Deduction features a number of subsidiary transcendental arguments. Each begins with a premise either about the self-attributability of mental items, *apperception*, or else a premise that affirms the necessity and universality of a feature of our experience of objects. Kant's strategy is to establish a specific theory of mental processing, *synthesis*, by arguing that its truth is a presupposition of and a necessary condition for the truth of such a premise, and then to show that the categories have an essential role in this mental processing. On a metaphysical idealist interpretation of his position, the

objects of experience are produced by this mental processing, and it is due to the role that the categories have in this production process that they legitimately apply to these objects.

For Kant the most significant rival theory of mental processing is that of his target, Hume. Hume agrees that a theory of experience will feature an account of the processing of mental items, but he denies that such an account should involve a priori concepts, and a fortiori that it issues in their applicability to experience. In his theory, *associationism*, our mental repertoire consists solely of *perceptions*, all of which are sensory items—the more vivid impressions, and their less vivid copies, the ideas, which function in imagination, memory, reasoning, and conceptualization (1748, §2). Association itself is the process by which these perceptions are related and ordered (1748, §3). An important characteristic of association is that it allows no resources other than the perceptions themselves. How perceptions are ordered is accounted for solely by facts about the perceptions themselves. Significantly, a subject not constituted solely of perceptions has no role in Hume's theory: the Humean subject is just a collection of perceptions (1739: I, IV, vi). These last two features make Hume's associationism a particularly economical theory, which results in a prima facie advantage over Kant's more complex view. Kant contends, however, that associationism cannot accommodate the compelling premises of the Transcendental Deduction, and this makes the case for synthesis by a priori concepts.

For Kant, synthesis is 'the act of putting different representations together, and grasping what is manifold in them in one cognition' (A77/B103); it is a process that 'gathers the elements for cognition, and unites them to form a certain content' (A78/B103). Synthesis takes multiple representations—in Kant's term, a 'manifold'—and connects them with one another to produce a single further representation with cognitive content (Kitcher 1990, 2011). This process employs concepts as ways of ordering representations. A claim crucial to the Transcendental Deduction is that it is the categories by means of which manifolds of representations are synthesized. Because the understanding of the subject is the source of the categories, and since it is also the faculty that generates synthesis, the subject plays an essential role in mental processing. It is important for Kant's theory that this subject is distinct from its states, and this is a further respect in which it differs from Hume's.

Here we will focus on the core part (§§16–20) of the Transcendental Deduction in the second edition of the *Critique of Pure Reason* (1787/1987), the B-Deduction. On my reading, in §§16–20 of the B-Deduction Kant employs a two-pronged strategy for defeating associationism and establishing synthesis, each of which is a transcendental argument. The first, contained in §16, is designed to demonstrate that association cannot account for an aspect of consciousness of the self that Kant refers to as the consciousness of its unity, and that such an account requires synthesis instead or in addition. This kind of transcendental argument he calls an *argument from above*, signifying that it begins with a premise about self-consciousness. Correlatively, §§17–20 features an *argument from below*, by which Kant intends to establish that synthesis by the categories is needed as a necessary condition for certain features of how we represent objects (the above/below terminology derives from A119).

The argument from above in §16 can be divided into two stages. The aim of the first is to establish the various components of *the principle of the necessary unity of apperception*. The second stage aims to show that synthesis is a necessary condition for the aforementioned aspect of self-consciousness, which this principle highlights. *Apperception* is

the apprehension of a mental state, a *representation* (*Vorstellung*) in Kant's terminology, as one's own; one might characterize it as the self-ascription or self-attribution of a mental state (Strawson 1966: 93–4). In Kant's conception, apperception of my representations has *necessary* unity in the sense that all of my representations *must* be grounded 'in pure apperception, that is, in the thoroughgoing *identity* of the self in all possible representations' (B131–2, emphasis mine). By this he means that:

> *(The principle of the necessary unity of apperception)* It must be the case that each of my representations is such that I can attribute it to my self, a subject which is the same for all of my self-attributions, which is distinct from its representations, and which can be conscious of its representations. (A116, B131–2, B134–5)

Kant initiates the first stage of the argument in §16 by claiming:

> It must be possible for the "I think" to accompany all my representations; for otherwise something would be represented in me which could not be thought at all, and that is equivalent to saying that the representation would be impossible, or at least would be nothing to me. (B131–2)

On one interpretation, the sense in which a representation would be impossible or nothing to me if it could not accompanied by the "I think" is simply that I could not then become conscious of it (Guyer 1987: 139–44). It is credible that for any representation of which I am conscious, I can attribute it to myself as subject, assuming my mental faculties are in working order, and if no controversial account of the nature of the subject is presupposed. But the claim that I can become conscious of *each* of my representations, and that it is therefore possible for me to attribute each of them to myself as their subject, is likely to be false. Plausibly, some of my representations are so thoroughly subconscious that I cannot attribute them to myself, while they are nevertheless mine due to the causal relations they bear to other representations and to actions that are paradigmatically mine. Fortunately, however, the premise that each of my representations is such that I can attribute it to myself is not crucial for the argument from above. Rather, the one Kant ultimately singles out:

> I am conscious of the identity of myself as the subject of different self-attributions of mental states

is significantly less committed, and also highly credible and resilient.

The argument from above crucially turns on the proposal that only a priori synthesis can explain *how I might represent the identity of my apperceptive consciousness* (B133) or *how I might represent the identity of the apperceiving subject* (B135) for different elements of the manifold of intuition to which I can attach the *I think*. The inadequacy Kant claims for "empirical consciousness," that is, for consciousness according to Humean psychological theory, is that "it is in itself dispersed and without relation to the identity of the subject" (B133). One idea expressed here is that Hume's theory does not have the resources needed to account for one's ability to attribute representations to one's self conceived as a subject that is both conscious of them and the same subject for each act of self-attribution. Humean theory can accommodate the view that apperceptive consciousness consists in perceptions that are intrinsically self-conscious, or else perceptions of perceptions. But intrinsically self-conscious perceptions would be distinct from one another, as would perceptions of

perceptions; and thus they too would be "dispersed" (B133), and have no common subject. Hume might propose to explain one's sense of the identity of the conscious subject of different self-attributions by the perceptions of perceptions being components of a single causally coherent bundle. Still, this bundle would not itself be conscious of perceptions. Consciousness of perceptions would instead be an intrinsic feature of individual self-conscious perceptions or a feature of individual perceptions of a perception. On Kant's proposal, by contrast, accounting for one's sense of the identity of the conscious subject of different self-attributions requires that this subject be distinct from its representations, a view that Hume rejects.

The second stage of the argument from above of §16 involves a further implication of the claim that "empirical consciousness, which accompanies different representations, is dispersed and without relation to the identity of the subject" (B133), that is, that Hume's theory lacks the resources to account for my *representation-relation* to the identity of the subject. His view cannot explain how I can "represent to myself the *identity of the consciousness in [i.e. throughout] these representations*" (B133). We might imagine several kinds of explanation for my representation of this identity. One candidate is that inner sense accounts for it. On this suggestion, the way I represent the sameness of the subject would be akin to how I commonly represent the identity over time of ordinary objects—by sensory apprehension of intrinsic properties, and noting that these intrinsic properties remain the same, or similar enough, over time. However, Kant and Hume concur that this is not the way I could represent the identity of the apperceiving subject, since they agree that by inner sense I cannot represent any intrinsic properties of such a subject. A second kind of explanation, which Kant endorses, is that I have an indirect way of representing this identity. This representation must instead depend on my apprehending a feature of my representations (Allison 1983: 142–4; Guyer 1977: 267, 1987: 133–9). The appropriate feature is a type of unity or ordering. Kant's idea is that if the representations I can attribute to myself possess a unity of the right kind, and if I am conscious of this unity, then I will be able to represent the apperceiving subject of any one of them as identical with that of any other. My representation of the identity of the subject comes about 'only in so far as I conjoin one representation with another, and am conscious of the synthesis of them' (B133).

This consciousness is plausibly interpreted as conscious awareness, not of the act or process of synthesis itself, but rather of the unity that is its outcome (Strawson 1966: 94–6; Dicker 2004: 133–4)). What sort of unity must I consciously recognize among my representations that would account for my representation of this identity? A credible proposal is that the unity consists in certain intimate ways in which representations in a single subject are typically related. Arguably, the essential feature of this unity is that a subject's representations be inferentially and causally integrated to a high degree, and in this respect they are unified in a way in which representations possessed by distinct subjects are not. When mental states fail to exhibit inferential and causal integration, as in the case of multiple-personality disorder, we have a tendency to posit distinct subjects, and we do not when such integration is present.

In Kant's view, the candidates for accounting for this kind of unity—or, less ambitiously, for my ability to recognize this sort of unity—are association and synthesis. At this point in the argument he seems to suppose that because Hume's psychological theory has already been ruled out, synthesis is the only remaining option. So for me to represent the identity of

the subject of different self-attributions, I must generate or at least recognize the right sort of unity among these representations, and synthesis must be recruited to account for this unity. Thus Kant contends that this combination 'is an affair of the understanding alone, which itself is nothing but the faculty of combining a priori' (B134–5). Since the understanding provides concepts for synthesis, and because for synthesis to be a priori is, at least in part, for it to employ a priori concepts, Kant is contending here that synthesis by means of a priori concepts is required to account for the unity in question.

Here is an austere representation of the structure of the argument so far:

(1) I am conscious of the identity of myself as the subject of different self-attributions of mental states. (premise)
(2) I am not directly conscious of the identity of this subject of different self-attributions. (premise).
(3) If (1) and (2) are true, then this consciousness of identity is accounted for indirectly by my consciousness of a particular kind of unity of my mental states. (premise)
(4) This consciousness of identity is accounted for indirectly by my consciousness of the particular kind of unity of my mental states. (1, 2, 3)
(5) If (4) is true, then my mental states indeed have the particular kind of unity. (premise)
(6) This particular kind of unity of my mental states cannot be accounted for by association. (5, premise)
(7) If (6) is true, then this unity of my mental states is accounted for by synthesis by a priori concepts. (premise)
(8) This unity of my mental states is accounted for by synthesis by a priori concepts. (6 and 7)

The structure of this part of the deduction as a transcendental argument is clear. Premise 1 is intended as a claim the metaphysical-concept skeptic will accept. The crucial necessary conditions, expressed by (3) and (7), are at root necessary conditions of the only possible explanation type. It's not especially plausible, however, that Kant has ruled out all of the competing explanations. At the same time, the argument would still have force against the skeptic's position if the necessary conditions these premises express are those of best explanation. The skepticism targeted by the Transcendental Deduction does not question conditions of this sort.

Paul Guyer forcefully argues that establishing synthesis by a priori concepts would require ruling out the alternative view that empirical information and concepts derived from experience are sufficient to account for the recognition of the unity at issue (Guyer 1987: 146–7). In particular, it remains open, given what Kant has shown, that this recognition requires only awareness of information derived from inner sense or introspective experience. At this juncture in the argument from above Kant does not take on the task of ruling out such a rival empiricist proposal, but he would need to do so to establish the need for synthesis by a priori concepts.

In the next phase of the Transcendental Deduction (§§17–20), an argument from below, this is exactly the task Kant takes on. In §18 he draws our attention to certain features of our representations of objects that, in his view, will serve to defeat associationism, the

empiricist's rival proposal, and establish a priori synthesis (Ameriks 1978; Pereboom 1995, 2009; Kitcher 2011). For Kant, a key characteristic of our representations of objects is their objective validity. For a representation to be objectively valid it must be a representation of an objective feature of reality, that is, a feature whose existence and nature is independent of how it is perceived (Guyer 1987:11–24). Kant contends that our objectively valid representations must in a sense be necessary and universal. However, the empirical unity of consciousness, which involves an ordering of representations produced by association, can only be non-universal, contingent, and hence merely subjectively valid, by contrast with the transcendental unity of apperception, which is or involves an ordering that is universal and necessary, and is therefore objectively valid. In Kant's conception, it is the fact that the transcendental unity of apperception is generated by synthesis by a priori concepts that allows it to yield an ordering that is universal, necessary, and objectively valid.

To illustrate and support these claims, Kant here invokes the example of the ordering of phenomena in time that has the key role in the discussion of the Second Analogy (cf. Guyer 1987: 87–90; Dicker 2004: 137–44). There he argues that our representations, considered independently of their content, are always successive. For example, when I view the front, sides, and back of a house when walking around it, and when I watch a boat float downstream, my representations of the individual parts and states occur successively. The objective phenomena represented by these successive representations, however, can be represented as either successive or as simultaneous—I represent the positions of the boat as successive, but the parts of the house as simultaneous. So despite the representations in each of these sequences being s*ubjectively successive*, I represent the parts of the house as *objectively simultaneous*, and the positions of the boat as *objectively successive*. How might we account for this difference in objectivity despite the similarity in subjectivity (Melnick 1973: 89)?

The important clue for answering this question is that these representations of objective simultaneity and succession are universal and necessary. On Kant's proposal, it is the universality and necessity of our representing the parts of the house as simultaneous that accounts for our representing them as objectively simultaneous, and the universality and necessity of our representing the positions of the boat as successive that accounts for our representing them as objectively successive. Association is inadequate for accounting for this objectivity because it is incapable of yielding such universality and necessity, a defect not shared by synthesis.

A first approximation of the import of 'universal' in the house example is:

> (U) Any human experience of the parts of the house is an experience of these parts as objectively simultaneous.

The addition of necessity has the following result:

> (U-N, first pass) Necessarily, any human experience of the parts of the house is an experience of these parts as objectively simultaneous.

Hume would resist this claim if the necessity were specified as ranging over all possible circumstances, since his theory would allow for the possibility of a deviant ordering in unusual empirical conditions. But (U-N, first pass) can be reformulated more precisely as

(U-N) Necessarily, if empirical conditions are normal, any human experience of the parts of the house is an experience of these parts as objectively simultaneous.

Kant's proposal is that, given only the resources of association, the truth of (U-N) cannot be explained. His reason is 'whether I can become empirically conscious of the manifold as simultaneous or as successive depends on circumstances or empirical conditions,' and so 'the empirical unity of consciousness, through association of representations, itself concerns an appearance, and is wholly contingent' (B139–40). Association does not yield an explanation the truth of (U-N), for given only the resources of association, the parts of the house will not necessarily or universally be represented as objectively simultaneous even supposing only normal empirical conditions. Kant has us consider an activity, word association, which functions as a paradigm for association. Word association, familiarly, does not yield universal and necessary patterns: 'one person connects the representation of a certain word with one thing, the other [person] with another thing.…' (B140). Hume's own paradigm for association in is the relations among parts of a conversation (1748: §3). In conversations people make different associations in similar circumstances. Kant's point is that if the paradigms for association fail to exhibit the sort of necessity and universality at issue, then the proposal that association can yield such an ordering of representations—wherever we might find it—is excluded.

Here we should see Kant as advancing his claim for the applicability of the categories by ruling out association as an explanation for (U-N). The structure of the resulting transcendental argument can be represented as follows:

9. We have representations of objects, i.e. of objectively valid phenomena. (premise)
10. All of our representations of objects are of universal and necessary features of experience. (premise)
11. Necessary and universal features of experience cannot be explained by association. (premise, from reflection on the nature of association)
12. If (10) and (11) are true, all of our representations of objects require a faculty for ordering mental states distinct from association. (premise)
13. All of our representations of objects require a faculty for ordering mental states distinct from association. (11, 12)
14. If (13) is true, all of our representations of objects require a faculty for synthesis by a priori concepts. (premise)
15. All of our representations of objects require a faculty for synthesis by a priori concepts—that is, the same faculty that accounts for my consciousness of the identity of myself as the subject of different self-attributions of mental states. (8, 13, 14)

To this we can add the final moves, which are explained in the subsequent sections of the B-Deduction:

16. Insofar as our representations of objects require a faculty for synthesis by a priori concepts, certain a priori concepts—the categories—legitimately apply to these objects. (premise)

C. We have representations of objects, and they are all such that the categories legitimately apply to these objects. (9, 15, 16)

The key necessary conditions expressed by (12) and (14), like those of the argument of the first stage, are conditions of only possible explanation. Here again, if these conditions turned out to involve best explanation instead, the argument would retain its force against the targeted skeptic.

In summary, the challenge Kant issues in this second stage of the Transcendental Deduction is to explain why, under normal conditions, ordering of representations in experience is universal and necessary. Part of the only explanation, he believes, is that we must have a faculty for ordering the representations. Hume might agree with this conclusion, supposing a sufficiently thin conception of 'faculty' on which it might consist solely of sensory items and associative tendencies among them. Kant argues that the Humean proposal cannot account for the truth of propositions such as (U-N), for the very paradigms of association, such as word association, and the association of topics in a conversation, do not exhibit the requisite universality and necessity. The alternative that can account for the truth of propositions such as (U-N) involves affirming the conclusion (C), that we have a faculty for synthesis by a priori concepts, which is the same faculty that was shown earlier to be required for my consciousness of the identity of myself as the subject of different self-attributions of mental states.

Briefly, here is the rest of the story. In §19, Kant argues that there must be a certain way in which each of my representations is unified in the subject, and he identifies this way with judgment: 'I find that a judgment is nothing but the manner in which given cognitions are brought to the objective unity of apperception' (B141). Judgment, Kant proposes, is objectively rather than subjectively valid, and hence exhibits the type of universality and necessity that characterizes objective validity (B142). He then claims that without synthesis and judgment as its vehicle, an ordering of representations might reflect what appears to be the case, but it would not explain how we make distinctions between objectively valid phenomena (i.e. objects) and the subjective states they induce. In §20 Kant ties this notion of judgment to the twelve forms of judgment presented in the Metaphysical Deduction (A70/B95), and then connects these forms of judgment to the twelve categories (A76–83/B102–9). The challenge has often been raised that the links Kant specifies between synthesis and judgment, judgment and the forms of judgment, the forms of judgment and the categories are not sufficiently supported (Guyer 1987: 94–102). Béatrice Longuenesse (1998), in her state-of-the-art interpretation of the Metaphysical Deduction, takes up this challenge with impressive results.

How resilient will the first premises for the two component transcendental arguments be? They, in effect, are:

(1) I am conscious of the identity of myself as subject of different self-attributions of mental states.

(9–10) We have representations of objects, all of which are of universal and necessary features of experience.

The Humean skeptic might try to reject (1), but denying this consciousness of subject-identity is a radical and unattractive move, even for Hume. Regarding (9–10), Hume would not disavow the necessity and universality this premise invokes, on a proper understanding

of the kind of necessity at issue. He maintains that it is in some sense impossible, given an experience of constant conjunction, that the mind not be carried from an impression of the first conjunct to an idea of the next:

> having found, in many instances, that any two kinds of objects, flame and heat, snow and cold, have always been conjoined together; if flame or snow be presented anew to the senses, the mind is carried by custom to expect heat or cold, and to *believe*, that such a quality does exist, and will discover itself upon a nearer approach. This belief is the necessary result of placing the mind in such circumstances. It is an operation of the soul, when we are so situated, as unavoidable as to feel the passion of love, when we receive benefits; or hatred, when we meet with injuries. (1748: §5)

Hume himself contends that given certain specific empirical circumstances, a particular type of ordering of perceptions in a sense necessarily (and universally) comes about, and this is just the type of claim Kant is making in (9–10).

The Transcendental Deduction has been highly influential as the paradigm of the method of transcendental argument. More specifically, it pioneers the alluring idea of using this method to draw significant anti-skeptical conclusions from premises about self-consciousness alone, and the now-standard tactic of arguing for concepts whose source is in the mind from universal and necessary features of experience.

3.2 The Refutation of Idealism

Kant's quarry in the Refutation of Idealism is Cartesian skepticism about the external world (B274–279; Bxxxix–Bxli). His intent is to refute what he calls *problematic idealism*, according to which the existence of objects outside of me in space is 'doubtful and indemonstrable' (B274). Kant's strategy is to show that the existence of such objects is a necessary condition of my awareness that my representations have a specific temporal order. At the present time I am aware of the specific temporal order of many of my past experiences. This awareness is produced by memory, but what is it about what I remember that allows me to determine the temporal order of these experiences? There must be something by reference to which I can correlate the remembered experiences that allows me to do this. However, first, I have no conscious states that can play this role. In addition, this reference cannot be time itself, for 'time by itself is not perceived'; as Guyer observes, it is not as if the content of memories of individual events are evidently indexed to particular times, the way in which sportscasts and videotapes often are (Guyer 1987). Kant argues that the only other candidate for this role is something outside of me in space, and something that is permanent (cf. First Analogy, B224–5).

Kant's proposal is perhaps made plausible by how we often actually determine the times at which our experiences occur. We use the observations of the sun's positions, or of the changing shadow on a sundial, or of a clock that indicates time by means of the period of a pendulum. Kant's argument can be viewed as exploiting this fact, together with the observation that there is no similar periodic process in our conscious experience considered independently of any spatial objects it might represent, and that we lack any awareness of time by itself, to show we must perceive objects in space. For then it would be only by reference to such objects that we can determine the objective temporal order of our experiences.

George Dicker sets out a compelling representation of Kant's argument (Dicker 2004, 2008):

(1) I am conscious of my own existence in time; that is, I am aware, and can be aware, that I have experiences that occur in a specific temporal order. (premise)
(2) I can be aware of having experiences that occur in a specific temporal order only if I perceive something permanent by reference to which I can determine their temporal order. (premise)
(3) No conscious state of my own can serve as the permanent entity by reference to which I can determine the temporal order of my experiences. (premise)
(4) Time itself cannot serve as this permanent entity by reference to which I can determine the temporal order of my experiences. (premise)
(5) If (2), (3), and (4), are true, then I can be aware of having experiences that occur in a specific temporal order only if I perceive persisting objects in space outside me by reference to which I can determine the temporal order of my experiences. (premise)
(C) I perceive persisting objects in space outside me by reference to which I can determine the temporal order of my experiences. (1–5)

Two of the most pressing objections that have been raised against the Refutation are that the skeptic would resist the first premise, and that the argument is vulnerable to an instance of Stroud's objection. So first, a skeptic could reject the initial premise on the ground of a general skepticism about memory (Allison 1983: 306–7). Bertrand Russell, for example, proposes that for all I know I was born five minutes ago (Russell 1912). On this skeptical hypothesis, I would be mistaken in my belief that I had experiences A, B, and C which occurred more than five minutes ago, first A, then B, and lastly C. It's credible that a skeptic who claimed that we lack adequate justification for a belief that external objects exist would also be disposed to contend that I lack justification for my belief that I had experiences that occurred in the past in that particular temporal order. Accordingly, Kant is not clearly justified in supposing that Premise (1) provides leverage against an external-world skeptic (cf. Dicker 2008, Chignell 2010).

Second, consider the proposal that states of the self are as well-suited as objects in space to function as a reference whereby I can accurately discern the temporal order of my past experiences. Imagine I had available as such a reference solely the mere appearance of a digital clock in one corner of my field of consciousness. This would not clearly be less effective than an actual clock in space (cf., van Cleve, reported in Dicker 2004: 207; Dicker 2004: 207). This objection is an instance of the type of concern Stroud raises against world-directed transcendental arguments, namely, that mere representation of some feature, by contrast with the existence of the external feature the skeptic targets, is all that can be established as a necessary condition of the first premise. To this one might reply, with Dicker, that there are in fact no states of the self that can serve as such a reference. However, and this is the deeper worry, according to Berkeley's idealist view in which the *esse* (to be) of objects in space is their *percipi* (to be perceived), any spatial object would amount to no more than mental states of the subject. But Berkeleyan spatial perceptions would seem to be as effective a reference by which to ascertain the temporal order of my past experiences as perceptions of objects distinct from my mental states (cf. Allison 1983: 300–1; Chignell 2010).

The Refutation of Idealism is an especially ambitious transcendental argument, and it has inspired many others for a similar conclusion. But critics largely agree that the Refutation itself falls to instances of certain standard forms of objection to transcendental arguments: that the skeptic need not commit to the initial premise, and that the argument can establish at most a conclusion about our representations or beliefs and not about mind-independent reality.

4. Contemporary Kantian Transcendental Arguments

4.1 Practical Transcendental Arguments.

Transcendental arguments against various sorts of skepticism were developed with vigor in the mid-twentieth century, and it was P. F. Strawson who led this effort. One of Strawson's most influential works is his essay on moral responsibility, 'Freedom and Resentment' (Strawson 1962). The reasoning in this article has not traditionally been interpreted as a transcendental argument, but recently Justin Coates (forthcoming) has made a strong case for such a reading. In Coates's account, the argument begins with the premise to which the moral responsibility skeptic would agree, that meaningful adult interpersonal relationships are possible for us. It continues by pointing out that relationships of this sort require that the participants show each other good will and respect, and that they be justified in expecting this of one another. Expectations for good will and respect in turn require susceptibility to the reactive attitudes, such as moral resentment, indignation, and gratitude, and in particular, justified expectations for good will and respect presuppose that the participants are apt targets of these reactive attitudes. But to be an apt target of the reactive attitudes is just what it is to be a morally responsible agent. Consequently, that we are morally responsible agents is a necessary condition of the possibility for us of meaningful adult interpersonal relationships.

Note that not all the connections among the steps of the argument are plausibly instances of appeals to logical or even metaphysical necessary conditions. True, some are: if being an apt target of the reactive attitudes is what it is to be a morally responsible agent, the necessary connection invoked would be conceptual or metaphysical. But if expectations for good will and respect do require susceptibility to the reactive attitudes, this would be plausibly a case of nomological necessitation, where the relevant laws are psychological. And given the sort of skepticism targeted, nomological necessitation is not too weak a connection; it is not called into question by the arguments of the moral responsibility skeptic.

Critics have in effect taken issue with a number of steps of this argument, for example that expectations for good will and respect require susceptibility to the reactive attitudes, and that justified expectations for good will and respect presuppose that the participants are apt recipients of the reactive attitudes. Perhaps human relationships do not require susceptibility to moral resentment and indignation, but only to the non-reactive attitudes of moral concern, disappointment, and sorrow (Pereboom 2001). Another avenue of criticism involves separating moral responsibility from being an apt target of the reactive attitudes.

It may be that a forward-looking, that is, what Strawson calls an 'optimistic' notion of responsibility, is all that's required for good relationships, and it is not characterized by being an apt target of the reactive attitudes. But these criticisms are controversial, and Strawson's argument is widely accepted and acclaimed.

Another prominent transcendental argument in the practical sphere is the sort Korsgaard (1996) develops for the claim that we must value ourselves as rational agents. Here is Robert Stern's (2012) representation of one such argument. It begins with a premise about rational choice, and crucially features the notion of one's practical identity, the distinctive nature of oneself as an agent, which may include, for example, being a Harvard philosophy professor and an American citizen:

1. To rationally choose to do X, you must take it that doing X is the rational thing to do.
2. Since there is no reason in itself to do X, you can take it that X is the rational thing to do only if you regard your practical identity as making X the rational thing to do.
3. You cannot regard your practical identity as making doing X the rational thing to do unless you can see some value in that practical identity.
4. You cannot see any value in any particular practical identity as such, but can regard it as valuable only because of the contribution it makes to giving you reasons and values by which to live.
5. You cannot see having a practical identity as valuable in this way unless you think your having a life containing reasons and values is important.
6. You cannot regard it as important that your life contain reasons and values unless you regard your leading a rationally structured life as valuable.
7. You cannot regard your leading a rationally structured life as valuable unless you value yourself qua rational agent.

C. Therefore, you must value yourself qua rational agent, if you are to make any rational choice.

Stern (2012) explains this argument as follows. To act is to do or choose something for a reason. But one has reasons to act only because of one's practical identity; one does not have reasons to act independently of that identity. However, a practical identity can yield such a reason only if one regards that identity as valuable. Merely being a father gives one no reason to care for one's children; rather, valuing one's fatherhood has this force. But one cannot regard a particular practical identity as valuable in itself—Korsgaard argues that this sort of realism about value is implausible. The only remaining explanation is that one regards it as valuable because of the contribution it makes to providing one with reasons and values by which to live. But then one must believe that it matters that one's life has the sort of rational structure that having such identities provides. However, to see that as mattering, one must regard leading a rationally structured life as valuable. Then, in conclusion, to regard leading such a life as valuable, one must see one's rational nature as valuable.

Various steps in this reasoning are again controversial, but this argument and others in the same family have attracted much attention, and it is a fine illustration of the potential of the methodology of transcendental argument.

4.2 Transcendental Arguments Against External World Skepticism

The second half of the twentieth century featured a revival of transcendental arguments against external world skepticism inspired by the example of Kant's Refutation of Idealism; (Strawson 1966; and see the essays in Stern 1998 and in Smith and Sullivan 2011). Perhaps the most prominent example is Strawson's main argument in *The Bounds of Sense* (1966), although Hilary Putnam's (1981) much-discussed argument from the causal theory of reference against this sort of skepticism has also been interpreted as a transcendental argument (Stern 2012). In *The Bounds of Sense* Strawson sets out a number of transcendental arguments inspired by Kant's Transcendental Deduction and his Refutation of Idealism. The one that is best known and most influential (1966: 97–104) is modeled on the Transcendental Deduction, but intentionally without invoking the controversial and arguably dated features of Kant's transcendental psychology. His target is a skeptic who claims that our experience consists just of sense-data, and thus does not feature objects 'conceived of us distinct from any particular states of awareness of them.' One might think of the skeptical target as a Berkeleyan account according to which the *esse* of spatial objects of experience is to be perceived (1966: 98).

The essential structure of Strawson's transcendental argument is as follows. It begins with the premise that every (human) experience is such that it is possible for its subject to become aware of it and ascribe it to herself. It is a necessary condition of the truth of this premise that in every experience the subject is capable of distinguishing a recognitional component not wholly absorbed by, and thus distinct from, the item recognized (1966: 100). To be capable of distinguishing these components it is necessary that the subject conceptualize her experiences in such a way as to contain the basis for a subjective component—how the experienced item seems to the subject—distinct from an objective component—how the item actually is. Strawson argues that 'collectively,' this comes to "the distinction between the subjective order and arrangement of a series of such experiences on the one hand, and the objective order and arrangement of the items of which they are the experiences on the other' (1966: 101). Conceptualizing experience as involving an objective order and arrangement of items amounts to making objectively valid judgments about it, which, in turn, requires the conclusion that experience must consist of a rule-governed connectedness of representations (1966: 98). Summarizing, from a premise about self-consciousness we can conclude, as a necessary condition, that the subject conceptualizes his experience so as to feature a distinction between 'the subjective route of his experience and the objective world through which it is a route,' where the experience of the objective world consists in a rule-governed order of representations (1966: 105).

As noted earlier, Barry Stroud, in his 1968 article 'Transcendental Arguments,' issued a telling objection to the enterprise of defeating the external-world skeptic by transcendental arguments of this sort. These arguments reason from some aspect of experience or knowledge to the claim that the contested feature of the external world in fact exists. In each case the existence of the external feature will not be a necessary condition of the aspect of experience or knowledge featured in the initial premise, because a belief about the external feature would always suffice. Although the claim about existence of the aspect of the external world could be secured if certain kinds of verificationism or idealism were

presupposed, these views are highly controversial. Moreover, one could make as much of an inroad against the skeptic armed with verificationism or idealism alone, without adducing the transcendental argument at all (Brueckner 1983, 1984).

Although Strawson's transcendental argument in *The Bounds of Sense* is not a specific target of Stroud's (1968), Anthony Brueckner (1983: 557–8) points out that it is susceptible to the line of criticism that Stroud develops. For Strawson's argument, despite its aim, can only conclude that experience must be conceptualized in a certain way, that is, such as to allow the subject to make the distinction between an objective world and her subjective path through it. This is not a conclusion about how a mind-independent world must be, but only about how it must be thought.

More recent development of world-directed transcendental arguments reflects chastened expectations about what they might establish. One more modest sort of transcendental argument begins with a premise about experience or knowledge that is acceptable to the skeptic in question, and then proceeds, not to the existence of some aspect of the external world, but in accord with Stroud's criticism, to a belief in the existence of some aspect of the external world. Stroud himself advocates a strategy of this sort (Stroud 1994, 1998), as does Stern (1998b). The kind Stern proposes begins with the premise that we think of the world as being independent of us, and it concludes, as a necessary condition of this premise, that we must think of it as containing enduring particulars. Such an argument does not claim that it is a necessary condition of this premise that there must exist such particulars. It contends only for 'a connection solely within our thought: if we think in certain ways, we must think in certain other ways' (Stern 1998b: 165). A belief or thought to which one reasons in this way would, in Stroud's assessment, have a certain *indispensibility*, 'because no belief that must be present in any conception or any set of beliefs about an independent world could be abandoned consistently with our conception of the world at all,' and it would be *invulnerable* 'in the special sense that it could not be found to be false consistently with its being found to be held by people (Stern 1998b: 166; Brueckner 1996).

Stern advances a conception of this modest sort of transcendental argument on which it targets a skeptic who questions whether certain beliefs cohere with others in one's set, by contrast with a skeptic who questions whether certain beliefs are true (Stern 1998b). A modest transcendental argument of this sort would aim to show that a belief whose coherence with the other beliefs is so challenged, coheres with them after all. The requisite coherence might be demonstrated by showing that the belief in question is actually a necessary condition of a belief that is indispensable (in some coherentist sense) to one's set. Mark Sacks (1998) objects that if at the same time one admits that the belief might not be true, one's sense that one is justified in holding the belief will be undermined. This worry seems serious. Sacks contends that it arises because of a tension between the coherentist theory of justification and the realist correspondence theory of truth that the external world skeptic presupposes. He points out that one might respond by accepting a coherence theory of truth as well, but this would be to adopt a version of idealism. Moreover, even if one accepted a coherence theory of truth, one would still have to admit that for specific instances of a belief one might be mistaken, even if one did think that one was justified in holding that belief on grounds of coherence.

5. Final Words

The legacy of Kant's Transcendental Deduction and Refutation of Idealism is the notion of a transcendental argument, which from an uncontroversial premise about our thought, knowledge, or experience reasons to a substantive and unobvious presupposition and necessary condition of this premise, often an anti-skeptical conclusion. Much of the effort spent devising transcendental arguments in the second half of the twentieth century focused on refuting skepticism about the external world, and the prospects for this project do not seem especially bright. But transcendental arguments can be recruited for other purposes, as indicated by Strawson's argument concerning moral responsibility and Korsgaard's argument about valuing oneself as a rational agent. It's credible that the reasons for pessimism about their significance for refuting external-world skepticism will not transfer to such other uses, and that therefore transcendental argument remains a promising philosophical methodology.

Bibliography

Allison, H. (1983). *Kant's Transcendental Idealism*. New Haven: Yale University Press; (2004), second edition.

Ameriks, K. (1978) "Kant's Transcendental Deduction as a Regressive Argument," Kant-Studien 69, pp. 273–87.

Bennett, J. (1966). *Kant's Analytic*. Cambridge: Cambridge University Press.

Berkeley, G. (1713). *Three Dialogues between Hylas and Philonous*. Robert Adams, ed., Indianapolis, Hackett, 1979.

Broughton, J. (2002) *Descartes's Method of Doubt*. Princeton: Princeton University Press.

Brueckner, A. (1983). "Transcendental Arguments I," *Noûs* 17, pp. 551–75.

Brueckner, A. (1984). "Transcendental Arguments II," *Noûs* 18, pp. 197–225.

Brueckner, A. (1996). "Modest Transcendental Arguments," *Philosophical Perspectives* 10, pp. 265–80.

Cassam, Q. (1999). *Self and World*. Oxford: Oxford University Press.

Chignell, A. (2010). "Causal Refutations of Idealism," *Philosophical Quarterly*, 60, pp. 487–507.

Coates, J. (forthcoming). "Responsibility Without (Panicky) Metaphysics."

Dicker, G. (2004). *Kant's Theory of Knowledge*. New York: Oxford University Press.

Dicker, G. (2008). "Kant's Refutation of Idealism," *Noûs* 42, pp. 80–108.

Forster, E., ed. (1989). *Kant's Transcendental Deductions*. Stanford: Stanford University Press.

Guyer, P. (1977). "Review of W. H. Walsh, *Kant and the Criticism of Metaphysics*," *Philosophical Review* 86, pp. 264–70.

Guyer, P. (1987). *Kant and the Claims of Knowledge*. Cambridge: Cambridge University Press.

Hume, D. (1739). *A Treatise of Human Nature*. Oxford: Oxford University Press, 1978.

Hume, D. (1748). *An Enquiry Concerning Human Understanding*. Oxford: Oxford University Press, 2005.

Kant, I. *Gesammelte Schriften*, ed. *Koniglichen Preussischen Academie der Wissenschaften*, 29 Vols. Berlin: Walter de Gruyter et al.

Kant, I. (1781/1787/1987) *Critique of Pure Reason* (trans. P. Guyer and A. Wood). Cambridge and New York: Cambridge University Press, 1997. (References are in the standard pagination of the 1st (A) and 2nd (B) editions. A reference to only one edition indicates that the passage appears only in that edition.)
Kitcher, P. (1990). *Kant's Transcendental Psychology*. New York: Oxford University Press.
Kitcher, P. (2011). *Kant's Thinker*. New York: Oxford University Press.
Korsgaard, C. (1996). *The Sources of Normativity*. Cambridge: Cambridge University Press.
Longuenesse, B. (1998). *Kant and the Capacity to Judge*. Princeton: Princeton University Press.
Melnick, A. (1973). *Kant's Analogies of Experience*. Chicago: University of Chicago Press.
Peacocke, C. (1989). *Transcendental Arguments in the Theory of Content*. Oxford: Oxford University Press.
Pereboom, D. (1990). "Kant on Justification in Transcendental Philosophy," *Synthèse* 85, 1990, pp. 25–54.
Pereboom, D. (1995). "Self-Understanding in Kant's Transcendental Deduction," *Synthese* 103, pp. 1–42.
Pereboom, D. (2001). *Living without Free Will*. Cambridge: Cambridge University Press.
Pereboom, D. (2009). "Kant's Transcendental Arguments," in *The Stanford Encyclopedia of Philosophy (Winter 2009 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2009/entries/kant-transcendental/>. Accessed September 23, 2015.
Putnam, H. (1981). *Reason, Truth, and History*. Cambridge: Cambridge University Press.
Russell, B. (1912). *The Problems of Philosophy*. London, Williams and Norgate; New York, Henry Holt and Company.
Sacks, M. (1998). "Transcendental Arguments and the Inference to Reality," in Stern (1998a), 67–83.
Smith, J. and P. Sullivan (2011), eds. *Transcendental Philosophy and Naturalism*. Oxford: Oxford University Press.
Stern, R. (1998a), ed. *Transcendental Arguments*. Oxford: Oxford University Press, 1998.
Stern, R. (1998b). "On Kant's Response to Hume: The Second Analogy as Transcendental Argument," in Stern (1998a), pp. 47–66.
Stern, R. (2012) "Transcendental Arguments," in *The Stanford Encyclopedia of Philosophy (Fall 2012 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2012/entries/transcendental-arguments/>. Accessed September 23, 2015.
Strawson, P. F. (1962). "Freedom and Resentment," *Proceedings of the British Academy* 48, pp. 1–25.
Strawson, P. F. (1966). *The Bounds of Sense: An Essay on Kant's Critique of Pure Reason*. London: Methuen.
Stroud, B. (1968). "Transcendental Arguments," *Journal of Philosophy* 65, pp. 241–56.
Stroud, B. (1994). "Kantian Argument, Conceptual Capacities, and Invulnerability," in *Kant and Contemporary Epistemology*, Paolo Parrini, ed., Dordrecht, Kluwer, pp. 231–51.
Stroud, B. (1998). "The Goal of Transcendental Arguments," in Stern (1998a), pp. 155–72.

PART IV

PHILOSOPHY AND ITS NEIGHBOURS

CHAPTER 24

PHYSICS AND METHOD

LAURA RUETSCHE

1. Prospectus

Philosophy of physics is not by any stretch of the imagination normal science. Its methodologies are numerous and unsettled. The first part of this contribution offers a list of some methodological inclinations at large in the community of philosophers of physics. The list is unsystematic and incomplete. Still it conveys some sense of the plethora of methodologies, self-conscious and otherwise, to be found at the interfaces of philosophy, mathematics, and physics. The second part of this contribution describes and defends a methodological inclination I have. The inclination is to understand and pursue the project of interpreting physical theories in a way that runs counter to what I take to be a methodological disposition prevalent among philosophers of physics. This disposition is toward "Naturalism": the view that the only respectable metaphysics is the metaphysics that makes the best sense of our best physics.

2. Questions of Methodological Inclination

Not all questions that follow are answered, even implicitly, by all philosophers of physics.

2.1 Is Physics a Surrogate for Intuitions or an Arena for Putting Them to the Test?

In traditional "pure" philosophy, the constraints (such as they are) on theorizing emanate from the imaginations and cognitive sensibilities of theorizers, present and past. Assessments of conceptual good taste, of "naturalness," of what's unmysterious and what's counterintuitive, inform and propel the philosophical agenda. Often these assessments are prompted by thought-experimental settings incompatible with empirical science as we

understand it. Here's a psycho-sociological hypothesis: some philosophers of physics are philosophers of physics because pure philosophy gives them vertigo. They either lack faith in or lack outright the imaginative and judgmental capacities that drive pure philosophy. The possibilities the vertiginous take to animate their philosophical concerns aren't purely imaginative ones consolidated by unanchored judgments of naturalness and the like. They are, rather, possibilities according to physical theory. To be sure, imagination and intuition inform the very non-trivial move from the equations of a physical theory to its possibility space. (Much philosophy of physics takes the form of making this move.) The point isn't that philosophy of physics is intuition- and judgment-free, but rather that it can be practiced in a way that at least appears to minimize the role of free-standing philosophical judgment.

A limiting case of this may be the well-entrenched methodology of "theorem-proving." A question, say about the compatibility of quantum correlational phenomena with local common causal models, is cast in terms of the mathematical machinery of a physical theory. In terms of the same machinery, an answer is conjectured. So expressed, the conjecture can admit of outright proof or outright refutation. The results thus established are secured by appeal, not to intuition, but to mathematics. Of course, what secures their philosophical relevance is another matter. Some questions are better than others; likewise some translations of questions into the language of mathematical physics. By its nature, much of the theorem-proving literature builds upon, expands, and refines existing results, and so adopts at the outset existing frameworks for analysis. Thus not every exercise in theorem proving involves the hard conceptual work of crafting from the material of mathematical physics a felicitous framework for confronting questions about causality, determinism, identity, and so on.

Not all philosophers of physics shrink from the responsibility to own their philosophical judgments. Many bring strong commitments about *how the physical world is* to their engagement with physical theory. For these thinkers, a key challenge is to accommodate the success of physical theories superficially inconsistent with their commitments. Rising to the challenge bears fruit both physical and philosophical. Consider, for example, work on the foundations of quantum theories informed by the idea that the fundamental business of a physical theory is to posit a "primitive ontology": "a description of matter in space and time" (Allori et al. 2008). A homily of the Copenhagen orthodoxy is that quantum mechanics cannot be understood to offer any such description; one influential contemporary take on quantum mechanics identifies its ontology with a wave function living in a many-dimensional configuration space alarmingly unlike space and time as we know them (see Albert and Ney, 2013); another takes the theory to describe a teeming multiplicity of worlds (Wallace 2012). Convinced that accounts of physics that abjure a primitive ontology fail in their most basic explanatory duties, advocates of primitive ontology articulate live alternatives (e.g. Bohmian mechanics, or stochastic collapse theories understood as describing spacetime "decorations") that can be taken to possess primitive ontologies, and argue that these alternatives are superior resources for purposes both physical and philosophical.

2.2 A Horror of Metaphysics or a Taste for it?

As with the first question, I suspect that differences in temperament as much as doctrine govern responses to this question. There are philosophers of physics who disparage or disdain metaphysics, period, or metaphysics conceived as an enterprise distinct from the philosophy of physics. (This theme will be expanded in the second part of this chapter.) There are others in whose hands philosophy of physics is metaphysics by other means, a proving ground for metaphysical tools, a source of surprising challenges to entrenched metaphysical doctrine or new ways of thinking about old philosophical puzzles. Some examples: Maudlin 2007, which (among many other things) appeals to contemporary gauge theories to argue that fundamental physical properties should not be understood in terms of universals; North 2013, which uses a notion beloved of metaphysicians (grounding) to sharpen a debate beloved by philosophers of space and time (substantivalism vs relationism); a phalanx of authors on the question of whether the Principle of the Identity of Indiscernables holds of so-called identical particles (see Caulton 2013); Albert 2003, whose consideration of the thermodynamic arrow of time concludes that epistemic sanity hinges on adopting the "Past Hypothesis" that the early state of the universe was one of exceedingly low entropy.

2.3 Should we Inherit Questions from Other Parts of Philosophy, or Develop Proprietary Ones?

Clearly we can do both. Physicists are fluent in a strikingly theoretical language; they use this language to make claims about knowledge and empirical justification; the theories they develop afford us ways to describe predicaments in which epistemic agents might find themselves. It follows that one can engage with questions from the philosophical tradition in a way that's informed by physical theories. The consequence is sometimes to reframe the question, or to enrich the menu of possible answers. Some examples of this genre include Field 1973 (reference), Wilson 2006, and Maddy 2007 (the nature of concepts), Malament 1984 (the possibility of time-travel), and Peterson 2011 (self-locating belief in a quantum world). Naturally, this method gets tremendous purchase on issues in the general philosophy of science, such as the natures of physical law (Maudlin 2007, Lange 2009, Roberts 2008), scientific explanation/inter-theoretic reduction (Batterman 2002, Belot 2005), and the fate of determinism (Earman 1986).

A complementary methodology is to find one's philosophical provocation in theories of physics, which have a way of throwing up their own peculiar foundational and interpretive problems. A very brief list might include: the quantum measurement problem (Albert 2009), quantum entanglement (Cushing and McMullin 1989; Clifton and Halvorson 1999), the cosmological horizon and flatness problems (Earman and Mosterin 1999), and the "reduction" of thermodynamics to statistical mechanics (Sklar 1995).

[1] I am grateful to David Wallace for this way of formulating the question.

Some of the problems just listed are the conscious focus of physicists. One might well ask of philosophers engaged with problems proprietary to physical theories: do they take themselves to be in the business of contributing to physics itself? Entangled with the answer to this question is the answer to another: Are the tools appropriate to philosophy of physics technical or conceptual?[1] They can be both, of course. But those hard at work articulating the hierarchy of independence conditions in algebraic QFT (Summers 1990) clearly favor technical tools, whereas those developing a decision-theoretic analysis of quantum probability (Deutsch 1999) clearly favor conceptual ones. Physicists forging new theories have been known to cite philosophical sources of inspiration (Rovelli 2004, Smolin 2002). I would urge the community of philosophers of physics to refrain from the project of disabusing working physicists of their philosophically "mistaken" inspirations. If a callow misreading of the absolute-substantival debate motivates a genuine advance in quantizing gravity, then (maybe not so ironically) careless philosophy stands to make a greater positive difference to actual physics than careful philosophy does.

……..

That qualification aside, my own view is that all of the methodological options we have alluded to, along with others unmentioned, are worth pursuing. But I have my own favorites. And another appropriate project for philosophy of physics is the adjudication of such methodological choices. The balance of this essay will be dedicated to just one such adjudication.

3. Against Naturalism

Because I was raised by logical empiricists, I believe that metaphysics, at least the kind that might be informed by science [disclaimer: I think there are, and should be, other kinds!], is methodology. That is, the frameworks we choose reflect the aims we have for them, and their capacity to promote those aims is the only (scientifically salient) grounds for correctness of choice. Because of the sorts of theories I think about, I harbor a suspicion of the idea (call it "Naturalism") that there exist scientific grounds for identifying *any* single unified metaphysics as *the* one that makes *the best* sense of science. And so I harbor a correlate affinity for a counter position that "the metaphysics" of modern science (that is, the framing commitments that promote its many pursuits) are various, *and healthily so*.

A variety of factors reinforce my sentiments, and will figure in what follows. All can be stated in term of Sellars' memorable (but, I believe, ultimately inapt) metaphor of "the scientific image" (1963). The first reinforcing factor can be stated in Sellars' own words: the scientific image is still under construction. We don't have any final physical theories; their involvement in fostering their own successors is one dimension we should consider when evaluating the theories we do have. Another reinforcing factor is an appreciation of the non-trivial role "interpretation"—the project of equipping a physical theory with content by characterizing and circumscribing the space of possibilities it allows—plays in constituting the scientific image. A third reinforcing factor is attention to a significant family of physical theories, including quantum field theories (QFTs), that bear interpreting,

but whose pointfulness would be blunted by equipping them with a single uniform interpretation—by, that is, immersing them in "the" metaphysics that (purportedly) best develops their scientific image.

The second part of this contribution will unfold as follows. In section 3.1, I will distinguish the "Naturalist" position from another I call *the locavore position*, paying attention to the ways the project of interpretation informs each. Section 3.2 will identify and champion a methodological commitment that, in the arena of interpretation, gives the Naturalist position a fighting chance to establish itself, but leaves room for the locavore to operate. Section 3.1.3 will draw from projects of interpreting QFT considerations in favor of the locavore.

3.1 Naturalism and a Rival

A position shared widely by philosophers of physics is one I'll (following many of its proponents) label "Naturalism." Wallace states it pithily:

> Naturalism: the thesis that we have no better guide to metaphysics than the successful practice of science. (Wallace 2012, 58)

Compare Ladyman and Ross on the "radically naturalistic metaphysics" of *Every Thing Must Go*:

> The *raison d'etre* of a useful metaphysics is to show how the separately developed and justified pieces of science (at a given time) can be fitted together to compose a unified world-view. . . . one metaphysical proposal is to be preferred to another to the extent that the first unifies more of current science in a more enlightening way. . . . "metaphysics" [should] refer to the articulation of a unified world-view derived from the details of scientific research. (2009, 45, 66)

And Maudlin in *The Metaphysics Within Physics:*

> Metaphysics is ontology. Ontology is the most generic study of what exists. Evidence for what exists, at least in the physical world, is provided solely by empirical research. Hence the proper object of most metaphysics is the careful analysis of our best scientific theories (and especially of fundamental physical theories) with the goal of determining what they imply about the constitution of the physical world. (2007, 104)

According to Naturalism as I understand it, we should seek a metaphysics in the form of an account of the way the world is, an account based upon our best scientific theories; the success of those theories licenses us to believe the account. Thus Naturalists can take their coda from Musgrave: "We should be realists about what the best metaphysical considerations dictate, where the best metaphysical considerations are those that have yielded the best physics" (1992, 691).

I think that there is a plausible alternative to Naturalism. Because I would rather not term it "Un-Naturalism," I will (with apologies to Arthur Fine) call it *the locavore position*, and develop it as a reading of the take on the Scientific Realism debate Fine markets under the heading of "the Natural Ontological Attitude" (or NOA) and promotes in his marvelous monograph *The Shaky Game* and other writings (Fine 1986, 1996). While I attribute

the locavore position to Fine, I think it has roots in Carnap, and another formidable contemporary proponent in van Fraassen. My strategy for explicating the position will be to update the food metaphor governing NOA from Fine's original " 'California natural'—no additives, please!" (Fine 1986b, 177), which is very 1980s. On my twenty-first century reading, *The Shaky Game* calls for a *locavore philosophy of physics* that is a clear alternative to Naturalism.

One way to bring the locavore alternative into focus is to consider how Fine might respond to a very standard criticism of his position. To set this criticism up, consider a NOA homily:

> It seems to me that both the realist and the antirealist must toe what I have been calling "the homely line." That is, they must both accept the certified results of science as on par with more homely and familiarly supported claims . . . Let us say, then, that both realist and antirealist accept the results of scientific investigation as "true," on par with more homely truths. What distinguishes realists from antirealists, then, is what they add on to the core position . . . when we contrast the realist and the antirealist in terms of what each want to add to the core position, a third alternative emerges---and an attractive one at that. It is the core position itself, *and all by itself.* (1986, 127–9)

NOA adopts the core position all by itself. What Realism is supposed to add is a gratuitously percussive analysis of truth.

> The realist adopts a standard, model-theoretic, correspondence theory of truth; where the model is just the definite world structure posited by realisms and where *correspondence is understood as a relation that reaches right out to touch the world*. (1986, 137; my italics)

By contrast, NOA "adopts a no-theory attitude toward the concept of truth" (1986, 9):

> When NOA counsels us to accept the results of science as true, I take it that we are to treat truth in the usual referential way, so that a sentence (or statement) is true just in case the entities referred to stand in the referred-to relations. (1986, 130)

What, one may ask, is the difference between the Realist's and the NOA's "core acceptance of the results of science"?

> My colleague, Charles Chastain, suggested what I think is the most graphic way of stating the answer—namely that what the realist adds on is a desk-thumping, foot-stamping shout of "Really!" (1986, 129)

This brings us to a standard Realist criticism of NOA. It can be very simply put: *accepting as true under a no-theory theory of truth is all the Realist ever proposed; table-thumping was never a constitutive element of the doctrine.* Because real Realists don't stomp their feet, the supposed distinction between Fine's position and theirs evaporates.

To combat the sour note struck by this standard criticism, I will attempt to develop a flavor base against which NOA emerges as not only distinctive but also delightful. The key ingredients emerge from further elaborations Fine offers of NOA (much of them from the 1996 Afterword to the second edition of *The Shaky Game*):

> That attitude, to let science stand on its own and to view it without the support of philosophical "isms," is what characterizes NOA. All the *isms* . . . involve global strategies, approaches to interpreting and providing a setting for science as a whole. (1986, 9)
>
> [NOA] allows for different and competing answers, or perhaps no answers, as the case demands. [1996, 173]
>
> [It offers] opposition to global accounts and . . . focuses on local, contingent practice. [1996, 174]
>
> [It] seeks not to end philosophy, but to enlarge and redirect what it stands for. [1996, 175]
>
> [What it stands against is] a bad case of misplaced generality. [1996, 179]
>
> [what it asks is] how much universality is actually required for understanding. [1996, 180]

Clearly, NOA is against global additives. But is it against them because they're *additives* (the "California natural" complaint about the table thumpers) or against them because they're *global* (the locavore's complaint)? These aren't exclusive options, and it's pretty clear that NOA exercises both. What I think helps balance the sour note is seeing both *that* and *why* the locavore's complaint is the more fundamental.

Realism is the grandaddy of all *isms*. One way to dramatize the locavore's complaint is to consider the stock argument for scientific realism, that (in the words of Hilary Putnam)

> it is the only philosophy that doesn't make the success of science a miracle. That terms in mature scientific theories typically refer . . ., that the theories accepted in a mature science are typically approximately true, that the same term can refer to the same thing even when it occurs in different theories—these statements are viewed not as necessary truths but as part of the only scientific explanation of the success of science. (1975, 73)

Assuming the shape of this line to be tolerably familiar, I will call this "the Miracles Argument," and work with the following simple version of it:

1. Theory *T* is successful. (That is, it predicts, explains, unifies, coheres with other elements of our scientific worldview, underwrites technological advances, serves as an effective platform for the evaluation and development of other theories, and generally exhibits the galaxy of virtues we expect from our best scientific theories.)
2. T 's truth is the best explanation of this success.

Therefore *T* is (at least approximately) true.

The Shaky Game develops powerful criticisms of this and other abductive or "explanationist" defenses of Realism. One is that only people who are already Realists believe in abduction. "Hence no support accrues to realism by showing that realism is a good hypothesis for explaining scientific practice. If we are open-minded about realism to begin with, then such a demonstration … merely begs the question that we have left open ('need we take good explanatory hypotheses as true?')" (1986, 115). Another is that anything the Realist's stolid and bully *truth* can explain, a more subtle and *soigné* commodity—truth instrumentalized—can explain as well, and at less risk.

But let us suspend *these* criticisms. Let us suppose (what the Standard Criticism of NOA surmises) that the Realist and NOA are working with the same conception of truth, and let us, for now, simply bracket concerns about the adequacy of abduction. We can still make

out *yet another* criticism of the Miracles Argument, a locavore criticism that highlights the implication of interpretive projects in Realist attitudes and the Naturalist agenda. Directed at the schematic theory *T*, the Miracles Argument abstracts away from concrete plots of ongoing science. These are exactly the plots it is the locavore impulse to cultivate. Adopting a satellite-level view of the scientific garden and amalgamating these plots into the aspecific placeholder *T*, the Miracles Argument obliterates the particularities, of context and method and aim, that make those plots the plots they are.

Now it's one thing to have a temperament that favors specificity over sweep. It's another to justify the preference. So we haven't stated the locavore's criticism yet, just described her temperament. Her criticism is this: abstracting away from ongoing scientific practices, many and varied, the Realist obscures potential weaknesses in the Miracles Argument. To get at these weaknesses, imagine a Realist who attempts to appease the locavore by *de-schematizing* the Miracles Argument, say by setting *T* equal to quantum mechanics (QM). The deschematized Miracles Argument looks like

1. QM is successful.
2. QM's truth is the best explanation of its success.[2]

Therefore QM is (at least) approximately true.

QM is arguably the most successful scientific theory in human history. So the first premise of the de-schematized Miracles Argument is extraordinarily well supported. The move through the second premise to the conclusion rests on articles of abductive faith that we are refraining from questioning. This delivers the Realist to the conclusion—and leaves her in a pickle. The pickle is *how to believe the conclusion*.

Here the Realist confronts what van Fraassen calls "*the question of interpretation*: Under what conditions is the theory true? What does it say the world is like?" (1991, 242). She confronts the question because she can't have a contentful belief in QM without understanding the theory, and "to understand a scientific theory, we need to see how the world could possibly be the way that the theory says it is. An interpretation tells us that" (1991, 336–7). The modal dimension of the interpretive project is worth emphasizing. Saying what the world is like (according to a theory) entails saying something about how that theory organizes its possibility space. It entails saying something about what physical magnitudes characterize and discriminate between possibilities recognized by the theory, and something about how those physical magnitudes are interwoven into lawlike relations. (This needn't require an account of physical law; a suitably perspicuous presentation of the theory's possibility space will do.) An interpretation of a theory is an account of the worlds possible according to the theory; such an account equips the theory with content. To understand a theory is to grasp

[2] One might suspect that the second premise is not de-schematized *enough* to be plausible. What's missing is an account of *how* QM's truth makes sense of its empirical success. Such a suspicion reinforces my claim that, in order to have any traction, the deschematized miracles argument needs to be accompanied by an interpretation of QM.

[3] Notice that all live physical theories fall into the hands of philosophers already partially interpreted. For instance, as 3.2. elaborates, quantum theories generally come equipped with constitutive commitments concerning canonical magnitudes and their algebraic interrelationships.

its content. To believe a theory is to accept that the actual world is among the worlds possible according to that theory. Belief in a theory is belief in a theory under an interpretation (or it is contentless).[3]

The locavore can already register one moral. It's that those who set out to interpret *particular* physical theories give content to positions in the Realism debate, even non-Realist ones. What Realists believe when they believe a theory T is an interpretation of T; various antirealisms about T correspond to various attitudes toward interpretations T admits. *This* moral, as it stands, doesn't on its own undermine the Miracles Argument. For it's consistent with the view that once advocates of clashing *isms* have decided what attitudes we ought to have toward empirically successful theories in general, local laborers will supply the interpretations of specific successful theories towards which we should take those attitudes. The moral is also consistent with the possibility that Realists, wielding the Miracles Argument, will emerge triumphant from the clash of *isms*.

Equipped with the notion of an interpretation of a physical theory, we now have a way to *describe* what the Realist does when she enacts the conclusion of the deschematized Miracles Argument. The Realist believes an interpretation of QM. But describing the task doesn't make it any easier. The problem is that the menu of interpretations on offer is, for more Realists, unappetizing in the extreme. Here's a quick rundown (see Albert 2009 for more):

- *Naïve Realism (QM_{NR})*
  Every physical magnitude pertaining to a quantum system has a determinate value. Pro: Under this interpretation, QM accords with our classically-tutored expectations, which is to say, it's unmysterious. Con: It's inconsistent when conjoined with a small number of extremely reasonable assumptions (Redhead 1987). Conjecture: any account of explanation, etc., that supports the miracles argument would endorse belief in the mere empirical adequacy of QM over belief in QM_{NR} + the negation of one of these assumptions. (This conjecture has a floating ad hominem counterpart: anybody who bought the miracles argument would blanche at Realism about QM_{NR} + the negation of one of these assumptions.)

- *Tractatus §7.* [QM_{T_7}]
  Only observables whose values can be predicted with certainty are determinate. Pro: Under this interpretation, QM is self-consistent. Con: if QM_{T_7} is true (in the sense that (i) it applies to physical systems which are also measuring apparatuses, and (ii) its dynamical law—Schrödinger evolution—holds universally), among the observables from which determinate values are withheld are those whose values (if they existed) would constitute measurement outcomes. If QM_{T_7} is true, measurements don't have outcomes. (This is one expression of the quantum measurement problem.) Since measurements have to have outcomes in order for there to be phenomena to save, if QM_{T_7} is true, it isn't empirically adequate. The Miracles Argument, which posits a theory's truth as the explanation of its empirical adequacy, fails spectacularly.

- *The textbook interpretation* [QM_{TI}]
  Only observables whose values can be predicted with certainty are determinate. During measurement, Schrödinger evolution is suspended in favor of wave function collapse, a *sui generis* time development that leaves systems in states enabling the prediction with certainty of the values of pointer observables. Pro: Under this interpretation, measurements have outcomes. Con: positing two incompatible sorts of state evolution, without offering a criterion for when one applies and when the other, QM_{TI} is at best incomplete and at worst inconsistent. In the worst case, Realism is a non-starter. In the best case, there's inductive

evidence that Realism is doomed. The difference between Schrödinger evolution and collapse is in principle empirically detectable; experiments have always upheld Schrödinger evolution; in-practice experimentally accessible regimes keep growing. So there's inductive evidence that, should QM_{TI} draw a line, experimentalists will cross that line and refute QM_{TI}.

- *Stochastic reduction* [QM_{SR}]
Only observables whose values can be predicted with (near) certainty are determinate. Schrödinger evolution approximates an underlying stochastic evolution tuned to mimic Schrödinger evolution for isolated systems and to mimic wave function collapse for large systems. Pro: Under this interpretation, QM is consistent with itself and with the fact that measurements have outcomes (recorded in something like positions) confirming standard quantum statistics. Con: isn't tuning a miracle? Also any unhedged version of QM_{SR} is liable to refutation as experiment advances.

- *The Bohm theory* [QM_{BT}]
Every system has a determinate position and a determinate velocity at every time; these quantities obey dynamical equations that guarantee that if it's ever the case that the distribution of positions among systems in an ensemble accords with the standard quantum statistics defined by some quantum state, then that distribution will always conform to the quantum statistics defined by the appropriate Schrödinger evolute of that state. Pro: Under this interpretation, positions are always determinate and we recover something (vaguely) like a classical particle ontology. Con: if QM_{BT} is true, relativity theory is false. Since the Miracles Argument appears to warrant Realism about each theory, the tasteful Realist has some explaining to do.

- *Many Worlds* [QM_{MW}]
My favorite version is somewhat old-fashioned: there is a continuous infinity of physical universes; the collection is described by a quantum state ψ that always undergoes Schrödinger evolution. Whenever a measurement occurs, each universe enjoys a determinate outcome; the quantum statistical algorithm applied to ψ gives a probability distribution over universes in the collection (i.e. the quantum probability ψ assigns outcome x is the measure of universes in the collection where outcome x occurs). Pro: cosmologists and Trekkies like this interpretation. Con: many details need to be worked out (see Wallace 2012 for a state of the art progress report). However they're worked out, the ontology may be unattractive to the Realist.

This menu isn't complete and its individual entries aren't developed as thoroughly as their advocates must like. Still, it is enough to suggest, I think accurately, that those who think they know how to believe the conclusion of the de-schematized Miracles Argument subscribe to interpretations of QM that mortify most Realists. The Many Worlds interpretation commits zany metaphysics; the Bohm theory implies the mere instrumental adequacy of a theory (special relativity) most Realists would much rather accept as true than QM; and so on, and so on.[4]

Considered schematically, the Miracles Argument for scientific Realism has real intuitive pull. But when it's deschematized and made concrete as an argument for realism about a particular, extraordinarily successful empirical theory, most realists don't know how, or can't bring themselves, to believe its conclusion. This confirms the locavore suspicion, that

[4] A sociological curiosity: There are two enclaves of those comfortable contemplating realism about QM: the New York/New Jersey enclave and the Oxford enclave. Each enclave has a view about what the realist's options are. The intersection of the enclaves' lists of live options is null. (In NY/NJ, it's stochastic reduction or Bohm; in Oxford it's Many Worlds.)

the Miracles Argument purchased its persuasive force only by abstracting away from the particularities of the scientific successes it invokes.

There are several ways a Miracles Argument run in abstraction could go wrong when made concrete. One is that the nature of the virtues constituting *T*'s success, invoked in premise 1 of the Miracles Argument, can't be decided in abstraction from *T*'s employment in ongoing science. Drawing attention to this possibility, the locavore emphasizes that in the Miracles Argument it's not just the "*T*" that's schematic but also the "successful." Prediction, explanation, strength, systematization, novel extendability, and so on, may be virtues of scientific theories, but (the locavore contends) they aren't virtues that admit a uniform explication whose details are independent of the details of the circumstances activating those virtues. Call this contention "the disunity of virtue." The disunity of virtue could derail a deschematized Miracles Argument because a virtue like "explanatory success" might look a lot more truth-conducive when considered as though unified and in abstraction, than it looks when considered as a virtue of a particular theory.

Any particular theory, the locavore urges, is put to a variety of uses and under a variety of circumstances. This suggests another possible hitch in bringing the Miracles Argument down to earth. There may be no single interpretation of a particular theory *T* under which all (or even enough) of the virtues cited in the first premise of the Miracles Argument pertain to the theory. Here the locavore contends that a successful scientific theory *T underdetermines its own interpretation*, and does so in such a way that different, and often incompatible, interpretations account for different parts of *T*'s total virtuousity. Call this contention "the disintegration of virtue." What derails the concretized Miracles Argument when virtue disintegrates is that no single interpretation possesses the totality (or enough) of the virtues cited in premise (1) as reasons to believe (given premise (2)) that *T* is true. To accept the conclusion of the concretized Miracles Argument is to believe *T under some interpretation*. The abductive basis for belief in *T* under an interpretation consists in those of *T*'s successes for which that interpretation can account. The locavore contends that the disintegration of virtue dramatically attenuates the abductive support for any contentful belief in *T*.

Notice, by the way, how the disintegration of virtue challenges another article of faith for some Realists: that underdetermination is the stuff of idle skepticism, remote "theoretical possibilities and baroque mathematical constructions" (Ladyman 2002, 231). Realists dismiss concrete examples of empirically equivalent theories as the result of "logico-semantic trickery" (Laudan and Leplin 1991, 463) or "cheap tricks" (Hoefer and Rosenberg 1994, 604). If candidates for belief are interpretations of our successful theories, and if our successful theories underdetermine their own credible interpretations, we don't need baroque mathematical constructions to see that the success of a theory fails to fix what we should believe about that theory. All we need is to notice that different, and rival, interpretations are required to make sense of different manifestations of theoretical success.

The difference between the Realist/Naturalist and the locavore can be expressed as a matter of operator ordering. Let *s* range over scientific successes credited to a theory *T*.[5] The *ism*s aspire to make sense of *T*. The Realist aims to do so by devising an account of how the world could possibly be that would explain why *T* succeeds as well as it does; this account

[5] We could just as well consider scientific success, period—the locavore's case would look even stronger!

forms the basis of the Naturalist's metaphysics. Such an account is an *interpretation* of *T*. Let *i* range over possible interpretations of *T*. Here's what the Realist and Naturalist hope is true, or true enough for the Miracles Argument to work:

(CYC) (∃i)(∀s) *i* makes sense of *s*.

Should the hope be fulfilled, the successes in the scope of the existential quantifier in (CYC) constitute reasons (according to the Miracles Argument) to believe the interpretation *i* that makes sense of them. By contrast, the locavore suspects that even the following is too much to ask[6]:

(LOC) (∀s)(∃i) *i* makes sense of *s*

Raising this suspicion, the locavore *does not* thereby despair of projects of making sense. Indeed, the locavore can be seen as commending that we *resituate* those projects, by starting from actual ongoing science, and in an attitude of attention to concerns animating its practice. The locavore's methodology is to pursue this project. The locavore's hypothesis is a prediction about the outcome of pursuing the locavore method: the prediction is that this project, properly conducted and executed, will eventuate in a range of sense-making strategies, different ones adapted to different circumstances—not in a *single i*, "the" ' scientific image, belief in which is sanctioned by the Miracles Argument. The locavore's complaint is that Realism and Naturalism proceeds as though science were a cyclops, but it isn't.

Now look back at pp. 470–1 dossier on NOA. So well does this dossier fits the locavore I've just described that I propose to understand NOA as (expressed by) locavore philosophy of science/physics. So understanding NOA suggests how one might contrast NOA with Realism in a way that survives Realists' contentions that all they ever meant by truth was exactly what Fine says we should mean. Even accepting that (and, for good measure, exonerating abductive arguments for Realism from the charge of question-begging), there is the cyclops distinction between the locavore and the Realist/Naturalist. The Realist believes the Cyclops hypothesis (CYC): there's some interpretation *i* s.t. for all (or for enough) scientific successes *s, i* makes sense of *s*. The locavore denies this. It's such a single *i* the Realist believes when she accepts science as true, and such a single *i* the Naturalist comprehends in her metaphysics. Whatever accepting such an *i* as true amounts to, the locavore denies that it can be done, because the locavore denies that there are appropriate candidates *i*s to do it to.

In 1986b, Fine describes NOA as "a pro-attitude."[7] There are several ways to hear this. The way of hearing that seems to have leaked into commentaries on NOA takes it to be a "a pro-attitude" in the sense that it is "for" science, like Superman is for truth, justice, and the American way. But remarking that Superman *stands for* those things points to another way NOA might be a pro-attitude. It might be a pro-attitude the way a pro-noun is a pronoun and a pro-seminar is a pro-seminar: what it stands for, that is, its content, varies with its context of deployment. If the content of the Natural Ontological Attitude depends on the circumstances under which we adopt the attitude, there is a follow-up question: how

[6] Recall "perhaps no answer at all;" see also Fine 1986, 170–1.

[7] I believe the homonymy with the Davidsonian term of art (1963) to be accidental.

should we understand/circumscribe the contexts that equip NOA with content? Both the question of how we ought to specify the content-fixing context, and the question of how we ought to constitute the attitude once the context is supplied, are meatily philosophical.

This way of thinking of things affords—*pace* the standard criticism—a sharp contrast between the locavore and Realists/Naturalists. For Realists like Richard Boyd and Naturalists like Ladyman and Ross, the context that would fix the content of our pro-attitude is "the mature sciences," more or less in their entirety. For the locavore, it is local bits of ongoing scientific practice—with different content-fixing strategies countenanced as appropriate to different contexts. Countenanced as well is the possibility of contexts to which *no* content-fixing strategies are appropriate.

3.2 A Methodological Distinction

As I've glossed NOA, it roots not in a no-theory theory of truth, nor in a hostility to philosophy, but in the suspicions expressed by the locavore's hypothesis that, once one pays actual science the attention it's due, the loose confederacy of local sense-making projects evinced by (LOC) might seem attainable, but the Cyclops hypothesis (CYC) certainly will not. I believe that thinking about quantum physics, even mathematically tractable versions thereof applied to purely theoretical problems, supports the suspicion. (So I also think that it's no accident that the progenitor of NOA is someone who has thought about quantum physics a lot.) In this section I will attempt to extract support (which I cheerfully admit falls dramatically short of proof) for the locavore's hypothesis from considerations of quantum field theory on curved spacetime.

3.2.1 *A Non-Uniqueness Problem*

In this section, I will try to say enough about where some quantum field theories (QFTs) come from to motivate the claim that the commitments by which we identify a physical theory—in the case of a quantum theory, I would contend these to be commitments to a constitutive set of canonical commutation relations—don't fix the content of that physical theory. They can leave open content-elaborating (that is, interpretive) questions of vital interest to physics itself. To make way for this claim, I will give a sketch of one way we come up with quantum theories (the *Hamiltonian Quantization Recipe*), in order to characterize a radical non-uniqueness that plagues QFTs.[8]

Some key difference between classical and quantum theories can be briefly stated. Consider classical mechanics for a particle of mass m moving in 1-dimension. The canonical magnitudes q (position) and p (momentum) coordinate a phase space of possible states. All other physical magnitudes are functions of q and p; for instance, the system's kinetic energy is given by $p^2/2m$. It follows that a system's classical state enables one to predict with certainty the values of all physical magnitudes pertaining to the system. These magnitudes don't form an undifferentiated heap. They stand to one another in stable relationships that

[8] See Ruetsche 2011, and references therein, for a more thorough treatment of issues discussed in this section.

reflect the theory's kinematic laws and symmetries. Arguably, standing to one another in these relationships is part of what makes the physical magnitudes the magnitudes they are. These relationships are encapsulated by an *algebraic structure* supplied by the *Poisson bracket*. Where physical magnitudes *f* and *g* are functions from the theory's phase space to the real numbers, their Poisson bracket *{f,g}* is given by

$$\{f,g\} = \partial f / \partial q \, \partial g / \partial p - \partial f / \partial p \, \partial g / \partial q$$

There is no need to be mesmerized by the symbols. An algebra is basically a way of forming linear combinations and products of magnitudes—an enterprise, the definition of kinetic energy just given should suggest, crucial to the business of physics. The Poisson brackets between canonical position and momentum observables assume a particularly simple form:

$$(\text{CPB}) \;\; \{p,p\} = \{q,q\} = 0; \{q,p\} = 1$$

For contrast, next consider a generic quantum theory. A (pure) quantum state corresponds to a vector ψ in a Hilbert space *H*. Physical magnitudes (aka *observables*) correspond to self-adjoint operators **A** on *H*. Their possible values are quantized. For each observable **A**, the state ψ determines a probability distribution over **A** 's possible values. Usually, these probabilities are different from *0* and *1*. And usually, there is a trade-off between ψ's capacity to predict **A**'s values and ψ's capacity to predict **B**'s values. The *commutator bracket*

$$[\mathbf{A},\mathbf{B}] = \mathbf{AB} - \mathbf{BA}$$

sets the terms of this trade-off. It also lends the collection of quantum observables an algebraic structure.

Thus quantum theories are unlike classical theories. But not absolutely unlike. Both sorts of theories come with their collections of physical magnitudes algebraically structured. And this fact is the germ of the Hamiltonian quantization recipe. Given a classical theory, one follows the recipe to obtain a quantum theory that is *the* quantization of that classical theory. The recipe is simply stated: to quantize a classical theory, *find a suitably quantum analog of its canonical Poisson bracket relations.* For example, to quantize the simple classical theory discussed above, we convert its canonical Poisson bracket relations (CPB) (which were:*{p,p}={q,q}=0;{q,p}=1*) to *canonical commutation relations*

$$(\text{CCR})[\mathbf{P},\mathbf{P}] = [\mathbf{Q},\mathbf{Q}] = 0; [\mathbf{Q},\mathbf{P}] = \mathbf{I}\,ih/2\pi$$

and (this is the hard part) we find a Hilbert space *H* and self-adjoint operators ***P*** and ***Q*** that act on *H* to satisfy these canonical commutation relations (CCRs). Such a triple <***P***, ***Q***, *H*> is known as a *representation of the CCRs*; the magnitudes **P** and **Q** furnishing the representation by satisfying the CCRs are known as *canonical magnitudes*. The recipe generalizes: more complicated classical theories have more complicated canonical Poisson brackets; still, to quantize such a theory, one promotes its canonical Poisson

brackets to canonical commutation relations, and finds a Hilbert space representation of those CCRs.

We do not rest there, of course. Interesting physics requires a variety of magnitudes woven into functional and nomic relationships. We procure this variety by using the canonical magnitudes to generate a richer set of interwoven magnitudes. More precisely, we expand our collection of magnitudes by forming linear combinations and products, and limits of sequences of linear combinations and products, of the canonical magnitudes. The resulting structured collection is *the algebra of quantum observables*. Under the standard understanding of how to take limits, it coincides with $\boldsymbol{B}(H)$, the set of bounded operators on the Hilbert space bearing the representation of the CCRs.[9]

Having constructed our observable algebra $\boldsymbol{B}(H)$, we're still not ready to do quantum physics. We need, as well as an account of what the physical magnitudes are, an account of what states—understood as valuations on these magnitudes—are physically possible. Given certain apparently innocuous assumptions about how good states behave (for which, see Redhead 1987), states on $\boldsymbol{B}(H)$ can be shown to correspond to $\boldsymbol{T}(H)$, the collection of density operators on H.[10]

The story so far: starting with a classical theory and following the Hamiltonian quantization recipe eventuates in a quantum theory whose observables reside in the observable algebra $\boldsymbol{B}(H)$ generated by a representation of the CCRs, and whose states are given by $\boldsymbol{T}(H)$, the collection of density operators on H. $\boldsymbol{T}(H)$ catalogs the possibilities the theory allows; $\boldsymbol{B}(H)$, and in particular the details of its descent from canonical observables yielding a representation of the CCRs, tells us how those possibilities are structured. The pair $<\boldsymbol{B}(H), \boldsymbol{T}(H)>$ is the germ of an interpretation of the quantum theory so-constructed.

At this point, a worrisome question emerges: can we, starting from the *same* classical theory and competently following the recipe, obtain *different* quantum theories? Useful recipes deliver consistent goods. If the outcome of the Hamiltonian quantization recipe is radically non-unique, it is at best an incomplete guide to the construction of serviceable quantum theories. Fortunately, the worrisome question has a standard, and reassuring answer. One component of the answer is the

> *Stone-von Neumann Theorem* (1931): Suppose T is a theory of classical mechanics whose degrees of freedom are finite in number. Then all Hilbert space representations of the CCRs arising from T are *unitarily equivalent*.

The other component is the claim that the formal relation of unitary equivalence mentioned in the theorem successfully explicates the interpretive notion of physical equivalence. To see the grounds for this claim, suppose Werner and Irwin find superficially disparate realizations of the CCRs arising from a classical theory T. Further suppose that

[9] Closed in the weak operator topology, $\boldsymbol{B}(H)$ is a von Neumann algebra. The physical magnitudes are given by its self-adjoint elements.

[10] The magnitudes that characterize physical situations, and the valuations on those magnitudes physical possibilities instantiate, constitute a *kinematics* for quantum theory. (We also need a dynamics, an account of how states/observables change over time. For the sake of brevity, I am confining attention to quantum kinematics.)

Werner's and Irwin's representations are unitarily equivalent. Then (and only then!) the quantum theories based on those representations are related as follows:

> There is an *algebraic structure-preserving* bijection from Irwin's collection of physical magnitudes to Werner's collection of physical magnitudes that maps Irwin's canonical magnitudes onto Werner's canonical magnitudes; and
>
> There is a bijection from the set of states Irwin regards as physically significant to the set of states Werner regards as physically significant; and
>
> These bijections "preserve empirical content": the predictions any Werner state makes about any set of Werner observables are exactly duplicated by the predictions the corresponding Irwin state makes about the set of corresponding Irwin observables, and vice versa.

When (and only when) Werner's and Irwin's representation are unitarily equivalent, Werner's quantum theory recognizes the *same set of physical possibilities* as Irwin's theory. When unitary equivalence fails, Werner recognizes physical possibilities without counterpart in Irwin's theory, and physical observables without correlate in Irwin's theory, and vice-versa. Werner and Irwin disagree not only about what's possible but also about which structures of physical magnitudes real possibilities instantiate. Put another way, they disagree not only about what's possible but also about what's physical.

Our working assumption is that the content of a physical theory consists in the possibilities it allows. This assumption licenses us to regard theories whose possibility spaces coincide as physically equivalent. Thus: quantum theories arising from unitarily equivalent representations of the same CCRs are physically equivalent. The Stone-von Neumann theorem alerts us that for suitable CCRs, all their representations are unitarily equivalent. This allays the uniqueness worry.

That is, until we reflect that sometimes the classical theories we set out to quantize involve *infinitely many* degrees of freedom. The paradigm example is classical field theory, which associates (e.g.) a field magnitude with each of the continuously many points of space. We can still carry out the Hamiltonian quantization recipe to quantize such theories. But the Stone-von Neumann theorem, presupposing that the theory to be quantized has only *finitely* many degrees of freedom, fails to apply to these quantizations. So when we set out to quantize a classical field theory, we can obtain (even continuously many) unitarily *inequivalent* Hilbert space representations of the CCRs encapsulating its quantization. Because the quantum theories arising from inequivalent representations are (presumptively) physically *inequivalent*, this is a disconcerting embarrassment of riches. The quantization recipe on its own generates nothing deserving the title "*the* quantized Klein-Gordon field." When it comes to QFTs obtained by Hamiltonian quantization, it seems that we don't really know what we're talking about.

3.2.2 *Toe-ing the Homely Line*

Recall the claim I set out to make plausible: sometimes, the commitments by which we identify a physical theory don't fix the content of that physical theory. Sticking to the example of QFT, those commitments are commitments to a set of CCRs. And they leave open such content-elaborating questions as: what representation of those CCRs should we work with? Should we even be working with Hilbert space representations at all? Which physical

magnitudes, beyond those furnishing a representation of the CCRs, are there? If Werner and Irwin are working within unitarily inequivalent representations of the CCRs for their QFT, their answers to many of these questions are outright rivals.

To bring this discussion back to the methodological questions distinguishing the locavore from the Realist/Naturalist, we need only register a claim about the application of physics that the examples in the next subsection will underscore. That claim is: Commitment to a theory *in a form that fosters that theory's capacity to function as physics* often requires settling answers to at least some of the content-elaborating questions. So, therefore, does toeing the homely line, of accepting science on its own terms. For in each circumstance, the content adequate to the scientific aims at hand are reasonable terms for a scientific theory to demand.

We can distinguish two strategies for deciding what to accept when we accept a scientific theory on its own terms:

- answer the entire slate of content-elaborating questions, thoroughly, antecedent to any applications of the theory; use the answer to sustain every application of the theory.
- adapt your answers, as well as the list of questions they're answers to, to the problem at hand.

The former strategy will appeal to those, like Realists and Naturalists, who are persuaded of the Cyclops hypothesis and susceptible to the Miracles Argument. The latter strategy is the locavore's. It has the methodological advantage over the Cyclops strategy that following the locavore's strategy, one could reach the position sought by the Cyclops strategy, if such a position is tenable. For it could be that the locavore's strategy throws up a collection of answers sufficiently uniform that they can be gathered under one cyclopsian interpretation that makes sense of all applications—the grail of the first strategy. But it could also be that the collection of locally generated answers exhibits no such uniformity. And this is the outcome the locavore hypothesis predicts. The locavore disagrees with the Realist/Naturalist about how locally vs globally acts of accepting science are specified, as well as about the status of the contention that widespread acts of local science acceptance can't be reduced to a single global act of science acceptance.

3.3 Supporting the Locavore

This concluding section draws from local projects of making sense of QFT on curved spacetime some consideration that seem to me to support the locavore.

"Fecundity"—the capacity to breed successful successor theories—is a scientific virtue. Indeed, it's a virtue central to prominent recent expressions of Realism. Boyd (1983) takes the wherewithal to successfully and efficiently *extend* science to be of the first importance. The "structure" about which (some) Structural Realists would be Realists is what's preserved over theory change (Worrall 1989). But even if we fix our attention on a single project of theory building, based on a single foundation, when we examine the actual pursuit of that project, we find that there isn't a single overarching program of content-elaboration

that underwrites that pursuit in all its variety. Take QFT in curved spacetime as a breeding ground for quantum gravity. The very same aspiration, pursued in different ways, legislates in favor of rival strategies of content-elaboration.

In classical gravity theory, Einstein's field equations related the distribution of matter and energy in a spacetime to its curvature, in whose terms the theory affords an understanding of gravitation effects. Semi-classical quantum gravity is a sort of halfway house from classical to quantum gravity. The idea of semi-classical quantum gravity is to introduce a quantum element to Einstein's field equations by replacing a classical commodity—the stress energy T_{mn}—with a quantum commodity—supposing the universe is suffused with a quantum field in state ϕ, the expectation value that state assigns the stress energy. Let $<T_{mn}>_\phi$ denote this quantum commodity.

The non-uniqueness of representations of the CCRs underlying a QFT galvanized a content-elaborating question: which states are physical for that QFT? For a QFT on curved spacetime harnessed to the project of semi-classical quantum gravity, the expression $<T_{mn}>$ affords some traction on this question. It turns out that the expression is in general ill-defined. The expectation value of stress energy—a quantity without which semi-classical quantum gravity is hamstrung—makes sense only for special representations of the QFT CCRs, called Hadamard representations, and the states defined on those representation. This suggests a strategy for interpreting QFT on curved spacetime in a way that secures that theory's fecundity-for-quantum gravity: regard only Hadamard representations as physical.

Now fostering semi-classical quantum gravity isn't the only way QFT on curved spacetime helps direct us toward future theories of quantum gravity. There is also the Hawking effect, according to which black holes are not really black, but radiate, slowly losing their mass until all that's left is a sea of Hawking radiation whose thermal signature reflects fundamental features of the evaporated black hole. The pattern of reflection is captured in the equations of black hole thermodynamics. Reproducing these equations, at least roughly, serves string theorists and others pursuing quantum gravity *ab initio* as something like a quasi-empirical constraint: if their theories accommodate the "phenomena" of black hole evaporation, they believe that they're on the right track. So QFT on curved spacetime supports work on quantum gravity by describing the quasi-phenomena of black hole evaporation that serve fledgling theories of quantum gravity as constraints.

Here's the rub: some features of Hawking radiation and black hole evaporation—to wit, how the thermal signature of the evaporating black hole registers with a distant observer—are best modeled by non-Hadamard representations of QFT on curved spacetime.

Take the content-elaborating question: what states are physically possible, according to QFT in curved spacetime? Insofar as semi-classical quantum gravity is cultivated for hints about future quantum theories of gravity, states must be Hadamard (that is, such as to yield an expectation value to the stress-energy observable). Insofar as the Hawking effect is pressed into service as the closest thing there is to "data" constraining quantum gravity, non-Hadamard states must be allowed to participate, for they are among those which articulate the phenomenology of Hawking radiation. The virtue of serving as a springboard for theory development is a virtue QFT in curved spacetime exhibits under different, and rival, strategies for content elaboration. Fecundity isn't a virtue with an essence delivered by a single overarching interpretation of QFT.

We can draw a similar moral from projects of Loop Quantum Gravity, string theory's rival in the search for final physics. The CCRs of Loop Quantum Gravity admit two classes of mutually unitarily inequivalent representations: continuous and "polymer" representations. The continuous representations work well for taking semi-classical limit (Fredenhangen and Reszewski 2006), which is one of the ways one shows one's theories of quantum gravity to be on the right track. The "polymer" representations are good for defining diffeomorphically-invariant states (Ashtekhar 2009). Preserving diffeomorphism invariance, and thereby preserving what is taken to be the central physical insight of Einstein's theory of gravity, is touted as the signal accomplishment Loop Quantum Gravity claims over other approaches to quantizing gravity.

Because continuous and polymer representations are unitarily inequivalent, they correspond to different strategies of content elaboration. (A signature of the difference: in polymer representations, there are physical magnitudes that admit a continuum of possible punctal value. In continuous representations, there are not.) Any interpretation[11] of Loop Quantum Gravity that accords it the virtue of having a semi-classical limit deprives it of the virtue of making sense of diffeomorphism invariance. These are both significant virtues, but there is no uniform interpretation of Loop Quantum Gravity that equips it with both.

These examples should sow some doubt over the status of the Cyclops hypothesis. But they shouldn't devalue locavoracious inquiry. Making contact with ongoing science and its aims is *how we find out* that QFT in curved spacetime *can* underwrite the project of semi-classical quantum gravity, and explicate the Hawking effect. It's *how we find out* that LQG has a semi-classical limit, as well as states invariant under diffeomorphisms. It's also *how we find out* that fecundity is a virtue served by different content-elaborating strategies in different contexts, and *how we find out* that genuine wage-earning scientific theories require a variety of content-elaborating strategies in order to discharge their scientific duties. These lessons advance causes both scientific and philosophical. One philosophical cause is figuring out how to go about understanding science. Here, I think, the locavore approach is a plausible and attractive alternative to Realism and Naturalism.

References

Albert, D. Z. (2003), *Time and Chance* (Cambridge, MA: Harvard University Press).

Albert, D. Z. (2009). *Quantum Mechanics and Experience* (Cambridge, MA: Harvard University Press).

Albert, D. Z. and Ney, A. (eds.) (2013), *The Wave Function: Essays on the Metaphysics of Quantum Physics* (Oxford: Oxford University Press).

Allori, V. et al. (2008). On the common structure of Bohmian mechanics and the Ghirardi-Rimini-Weber Theory, *The British Journal for the Philosophy of Science 59*: 353–89.

Ashtekar, A. (2009). Some surprising implications of background independence in canonical quantum gravity, *General Relativity and Gravitation*, *41*(9): 1927–43.

[11] Confession: this contribution assimilates "interpretation" to "choice of privileged Hilbert space representation." The assimilation oversimplifies, but it is beyond the scope of this essay to cancel the oversimplification.

Batterman, R. W. (2002). *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence*. (Oxford: Oxford University Press).
Belot, G. (2005). Whose Devil? Which Details?. *Philosophy of Science*, *72*(1): 128–53.
Boyd, R. N. (1983). On the current status of the issue of scientific realism. In Carl G. Hempel, Hilary Putnam & Wilhelm K. Essler (eds.), *Methodology, Epistemology, and Philosophy of Science* (pp. 45–90). Springer Netherlands.
Caulton, A. (2013). Discerning "Indistinguishable" Quantum Systems, *Philosophy of Science 80*: 49–72.
Clifton, R. & Halvorson, H. (1999). Bipartite-mixed-states of infinite-dimensional systems are generically nonseparable, *Physical Review A*, *61*(1): 012108.
Cushing, J. T. and McMullin, E. (1989). *Philosophical Consequences of Quantum Theory*. (Notre Dame, IN: University of Notre Dame Press).
Davidson, D. (1963). Actions, reasons, and causes. *The Journal of Philosophy*, *60*(23): 685–700.
Deutsch, D. (1999). Quantum theory of probability and decisions, *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, *455*: 3129–37.
Earman, J. (1986). *A Primer on Determinism*. (Heidelberg: Springer).
Earman, J. and Mosterin, J. (1999). A critical look at inflationary cosmology, *Philosophy of Science 66*(1): 1–49.
Field, H. (1973). Theory change and the indeterminacy of reference, *The Journal of Philosophy 70*: 462–81.
Fine, A. (1986 [1996]). *The Shaky Game: Einstein, Realism, and the Quantum Theory*. [2nd ed.] (Chicago: University of Chicago Press).
Fine, A. (1986b). Unnatural attitudes: Realist and instrumentalist attachments to science, *Mind*, *95*(378): 149–79.
Fredenhagen, K., & Reszewski, F. (2006). Polymer state approximation of Schrödinger wave-functions, *Classical and Quantum Gravity*, *23*(22): 6577.
Hoefer, C. & Rosenberg, A. (1994). Empirical equivalence, underdetermination, and systems of the world, *Philosophy of Science*, *61*: 592–607.
Ladyman, J. (2002). *Understanding Philosophy of Science*. (London: Routledge).
Ladyman, J. and Ross, D. (2007), *Every Thing Must Go*. (Oxford: Oxford University Press).
Lange, M. (2009). *Laws and Lawmakers: Science, Metaphysics, and the Laws of Nature*. (Oxford: Oxford University Press).
Laudan, L. and Leplin, J. (1991). Empirical equivalence and underdetermination, *The Journal of Philosophy*, *88*(9): 449–72.
Maddy, P. (2007). *Second Philosophy: A Naturalistic Method* (Oxford: Oxford University Press).
Malament, D. B. (1984). 'Time travel' in the Gödel Universe, *Proceedings of the Biennial Meeting of the Philosophy of Science Association*, Vol. 2: 91–100.
Maudlin, Tim (2007). *The Metaphysics Within Physics* (Oxford; New York: Oxford University Press).
Musgrave, A. (1992). Realism about what?. *Philosophy of Science 59*: 691–7.
North, J. (2013), "The Structure of Spacetime: A New Approach to the Spacetime Ontology Debate," ms.
Peterson, D. (2011). Qeauty and the books: A response to Lewis's quantum sleeping beauty problem, *Synthese 181*: 367–74.
Putnam, H. (1975). *Mind, Language and Reality: Philosophical Papers, ii*. (Cambridge: Cambridge University Press).

Redhead, M. (1987). *Incompleteness, Nonlocality, and Realism: A Prolegomenon to the Philosophy of Quantum Mechanics*. (Oxford: Clarendon Press).
Roberts, J. T. (2008). *The Law-Governed Universe*. (Oxford: Oxford University Press).
Rovelli, C. (2004). *Quantum Gravity*. (Cambridge: Cambridge University Press).
Ruetsche, L. (2011). *Interpreting Quantum Theories*. (Oxford: Oxford University Press).
Sellars, W. (1963). Philosophy and the scientific image of man, *Science, perception and reality*, *2*: 35–78.
Sklar, L. (1995). *Physics and Chance: Philosophical Issues in the Foundations of Statistical Mechanics*. (Cambridge: Cambridge University Press).
Smolin, L. (2002). *Three Roads to Quantum Gravity*. (New York: Basic Books).
Summers, S. J. (1990). On the independence of local algebras in quantum field theory, *Reviews in Mathematical Physics*, *2*: 201–47.
Van Fraassen, B. C. (1991). *Quantum Mechanics: An Empiricist View*. (Oxford: Clarendon Press).
Wallace, D. (2012), *The Emergent Multiverse: Quantum Theory according to the Everett Interpretation*. (Oxford: Oxford University Press).
Wilson, M. (2006), *Wandering Significance*. (Oxford: Clarendon Press).
Worrall, J. (1989). Structural realism: The best of both worlds? *Dialectica*, *43*(1–2): 99–124.

CHAPTER 25

LINGUISTIC AND PHILOSOPHICAL METHODOLOGY

PETER LUDLOW

1. Introduction

For the past half century there has been significant interaction between philosophers and linguists and this has led to the borrowing of prima facie philosophical methodologies by linguistics, and correlatively, the borrowing of prima facie linguistic methodologies by philosophy. Both directions of methodological borrowing have led to misunderstandings—particularly in the philosophical community. In this essay my goal is to give some examples of this methodological borrowing, clarify what the aim of it has been, and ultimately make the case for the fruitfulness of these efforts.

I'll begin with the use of philosophical methodology in linguistics and then I'll take up the role of linguistic methodology in philosophy—particularly the role of appeals to semantics in metaphysics, epistemology, and the philosophy of language. As we will see, the appeal to semantics in philosophy has met with pushback, but as I'll argue this is largely driven by a misunderstanding about the nature of semantics. Semantics, I'll argue, is a theory of language/world connections, and the basic entities that semanticists deploy (including referential contents) are the product of investigation (scientific and otherwise) of the external world. Since semantics is concerned with the language of every aspect of human inquiry it follows that semantics includes our most complete theory of the world. Dismissing semantic claims because they have no purchase in (for example) physics may overlook the need for semantic properties that are needed elsewhere—for example in the theory of human action.

We'll also see that even at the level of particular methods, the exchange between philosophy and linguistics has never been unidirectional: philosophical methods introduced into linguistics are routinely recycled back into philosophy. Accordingly, I raise the question as to whether our prima facie distinction between linguistic and philosophical methodology even holds up in any interesting sense. I'll suggest that it doesn't.

2. Philosophical Methodology in Linguistics

We have all read screeds by scientists and academic pundits to the effect that philosophy is not particularly relevant to scientists. One recent example that caught the attention of the philosophical community[1] was a book review by Freeman Dyson in the *New York Review of Books*.[2]

> The fading of philosophy came to my attention in 1979, when I was involved in the planning of a conference to celebrate the hundredth birthday of Einstein. The conference was held in Princeton, where Einstein had lived, and our largest meeting hall was too small for all the people who wanted to come. A committee was set up to decide who should be invited. When the membership of the committee was announced, there were loud protests from people who were excluded. After acrimonious discussions, we agreed to have three committees, each empowered to invite one third of the participants. One committee was for scientists, one for historians of science, and one for philosophers of science.
>
> After the three committees had made their selections, we had three lists of names of people to be invited. I looked at the lists of names and was immediately struck by their disconnection. With a few exceptions, I knew personally all the people on the science list. On the history list, I knew the names, but I did not know the people personally. On the philosophy list, I did not even know the names.
>
> In earlier centuries, scientists and historians and philosophers would have known one another . . . But in the twentieth century, science and history and philosophy had become separate cultures. We were three groups of specialists, living in separate communities and rarely speaking to each other.

I can't speak for the relevance of philosophy to physics,[3] but Dyson at a minimum overgeneralizes when he extends his critique of the relevance of philosophy to science in general: the relation between linguistics and philosophy is a clear counterexample.[4]

[1] See discussion on Leiter Reports: A Philosophy Blog. <http://leiterreports.typepad.com/blog/2012/10/why-the-new-york-review-of-each-others-books-asked-freeman-dyson-to-review.html>. Accessed September 23, 2015.

[2] See Dyson (2012).

[3] Of course Dysons' ignorance of people working in the philosophy of physics may say more about him than about the philosophy of physics. How one can comment on the state of a field when one doesn't know the players much less what they have said strikes me as puzzling. Of course his point may simply be about the *impact* of philosophy in science and thus his argument may extrude from the assumption that if he doesn't know the players they can't be relevant. I consider this an exceptionally thin critique.

[4] In this paper my focus is on the relation between linguistics and philosophy but the relevance of philosophy to cognitive science is surely significant as well; arguably figures like Fodor, Chomsky, Dennett, Stich, Block, Goldman, and many others were important figures in the construction of cognitive sciences as we know it. Or to take another example, consider Arthur Burks, who for many years was a professor in the philosophy department at The University of Michigan. Not only did he play a role in the development of the UNIAC computer, but he also helped found the computer science department at Michigan. The general rule seems to be that philosophers continue to play important roles in emerging sciences, whether or not they play significant roles in mature sciences. If it is the case that the contribution of philosophy to mature science is minimal I would guess that the real

Linguistics—in particular generative linguistics—is the study of the human language faculty, and it attempts to explain the mechanisms and principles that give rise to human linguistic competence. Over the past half-century communication and collaboration between philosophers and linguists has been extraordinarily robust. Many philosophers have held joint appointments in linguistics (e.g. Paul Pietroski, Richmond Thomason, and myself), some philosophers have held full-time positions in linguistics (e.g. James Higginbotham), others have moved permanently into linguistics (e.g. Norbert Hornstein), linguists have been housed in philosophy departments (e.g. Tom Wasow and Bob Frieden), and some linguists have moved into philosophy (Robert May being an example). Dry appointments (positions in which philosophers and linguists are listed as affiliated members of other departments) are so commonplace that they number in the hundreds. Joint publications by philosophers and linguists may well number in the hundreds as well.

These interactions and collaborative publications have not been merely at the periphery of philosophy and linguistics: in many cases linguists have contributed to core problems in philosophy and philosophers have contributed to core issues in linguistics. On the latter point, it is really hard to imagine that generative linguistics would exist in anything like its current form without the contribution of philosophers.

Chomsky, who surely counts as a philosopher in his own right, has reported (1975; introduction) being influenced by his interactions with Nelson Goodman and Henry Hiz, and his early formal tools were borrowed from the logician Emile Post, who is arguably an honorary analytic philosopher. Throughout his career Chomsky has engaged with the leading figures of analytic philosophy, from Quine and Davidson, to Dummett, Burge, and Searle. Quite apart from the philosophical dispositions of Chomsky, as we will see, the development of generative grammar was heavily influenced by a number of philosophers.

For example, the symbiotic relationship between linguistics and philosophy was evident in the 1960s when work by Katz and Fodor (1963) and Katz and Postal (1964) laid the foundations for what would become known as Generative Semantics—a research program within generative linguistics on which many linguistic forms were transformationally derived from "deep structure" representations that by hypothesis encoded the meanings of natural language sentences. So, for example, a passive form like "The burger was eaten by John" was derived from a deep structure representation like "John ate the burger". A key idea articulated by Katz (a philosopher) and Postal (a linguist) was that there was a one–one isomorphism between meanings and Deep Structure representations, suggesting, for example, that synonyms had a transformational relationship to each other, or for example, that a verb like "kill" could be derived from "cause to die".

The Generative Semantics theory eventually collapsed (see Newmeyer 1986 for discussion), but a number of important linguistic analyses survived the collapse of the Generative Semantics research program (see, for example, den Dikken et al. 1996, Larson et al. 1997). It is also important to understand that the collapse of Generative Semantics was driven

problem with philosophy's ability to contribute to physics and other established sciences is simply a lack of bodies to throw at these subject matters. Philosophy departments are dwarfed by the sciences, and the charter of philosophy departments is to engage in critical thinking in areas that range from the foundations of cognitive sciences and linguistics, the philosophy of mathematics, the philosophy of logic, aesthetics, applied ethics and meta-ethics, and the list could go on and on. There just aren't enough philosophers.

in no small measure by arguments provided by philosophers. For example, Lewis (1972) argued that the basic philosophical position underlying the Katz-Postal proposal was suspect because it did not properly anchor meanings to the world. Meanwhile Fodor (1970) offered reasons for thinking that "kill" wasn't derived from "cause to die" (the short form of the argument: "cause to die tomorrow" is ambiguous in a way that "kill tomorrow" isn't).

In the 1970s it became apparent to a number of linguists that a more productive account of the semantics of natural language and the theory of meaning could be obtained by utilizing a model-theoretic semantics in the spirit of Montague (1974) or alternatively a truth-conditional approach in the spirit of Davidson's (1967a) "Truth and Meaning." The former approach was incorporated in, for example, textbooks by Dowty, Wall, and Peters (1981), and Heim and Kratzer (1998) and the Davidsonian approach in a text by Larson and Segal (1995).

The Generative Semantics period (and the immediate aftermath) was not the last period of contact between philosophy and generative linguistics. The 1970s saw the introduction of a large number of philosophical resources into linguistic theory. For example, in a development of generative linguistics now known as "The Extended Standard Theory," Chomsky (1977) proposed treating the extraction site of a moved noun phrase as being a "trace"[5] that would be bound either by the moved WH word or the quantified noun phrase that moved out of the position.

To illustrate, in example (0), the WH word has moved from the position where it is generated and is coindexed with the trace that it left behind.

(0) Who_i did John see e_i

The proposal (which was subsequently developed in Wasow (1972) and Chomsky (1977)) was larded with philosophical ideas.

The introduction of trace (and thus operators and variables) in the syntax played a role in a number of very fruitful syntactic constraints in linguistic theory. Two examples of these constraints are the role of trace in the explanation of "weak crossover" and movement asymmetry.

Crossover facts (initially noticed in Postal (1971)) can be illustrated by example (1), where "who" and the pronoun can be understood as co-referential while in (2) this is not possible.

(1) Who said Mary kissed him?
(2) Who did he say Mary kissed?

Wasow (1972) noted that these facts could be explained by trace theory. Specifically, when we look at these structures with trace introduced, we get structures like the following, where the 'e' is the trace left behind by the movement of the WH expression (in this case 'who').

[5] We can think of a trace as being a syntactic object that remains in the position from which the operator is extracted. It is typically represented with either a 't' or an 'e' which is subscripted with an index (for example 'i') that is shared with the moved operator. In this way the operator is indexed to its site of extraction.

(1t) Who_i [e_i said Mary kissed him_i]
(2t) *Who_i [he_i say Mary kissed e_i]

The grammatical constraint that blocks this sort of binding relation came to be called the the "leftness condition" (Chomsky 1976): A pronoun cannot be coindexed with a trace to its right (as in (2t)).

The introduction of trace not only helped to account for binding facts, but Fiengo (1977) argued that a large class of movement asymmetries could be accounted for if we thought of traces as behaving like bound anaphors, where an example of a bound variable would be a reflexive as in (3). To see how this works contrast behavior of reflexives in cases like (3) and (4).

(3) John likes himself
(4) *Himself likes John

To explain this, linguists have introduced a notion of syntactic scope, which they call "c-command."[6] The contrast between these cases is explained by the fact that in (3) 'John' c-commands (has syntactic scope over) 'himself' while in (4) it does not.

If traces, like reflexives must be c-commanded by (in the scope of) their antecedents, then any movement that involves moving an element lower in a clause would violate this principle, because if it went lower it would no longer c-command its trace. Chomsky (1973) had earlier gone out on a limb in suggesting that all the early transformations of the 1960s could be swept away in favor of a single rule stating that one can move anything anywhere and the introduction of a handful of constraints on movement. Fiengo's observation helped vindicate this proposal by showing that with the introduction of trace and pre-existing principles governing bound variables one can immediately rule out an entire class of movements—that is, downward movement.

Another interesting feature of the Extended Standard Theory was the introduction of the level of representation LF (suggesting a similarity to the philosopher's notion of logical form). The idea was that LF could be thought of as the syntactic level that interfaced with the semantics. The level LF involved a rule mapping from surface structure (SS) to LF. Called QR, the rule simply said "adjoin quantified NP to S." (May 1977). So there would be a mapping from a level of syntactic representation called "D-structure" to a more surfacey level of representation called "S-structure," and from S-structure to a level of representation that could be thought of as the input to the meaning system called "LF." There would also be a mapping from S-structure to PF or phonetic form yielding the following T-model of the grammar.

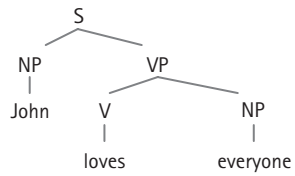
$$\begin{array}{c} \text{DS} \Rightarrow \text{SS} \Rightarrow \text{LF} \\ \Downarrow \\ \text{PF} \end{array}$$

[6] It can be defined off of phrase markers in the following way: node A c-commands node B just in case the first branching node that dominates A also dominates B?

To see how quantifier raising worked, consider sentence (5) and its SS representation (5-SS).

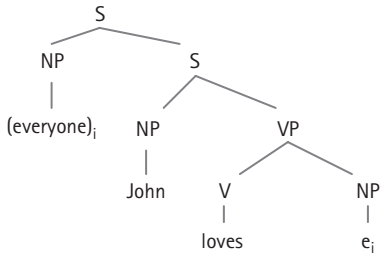
(5) John loves everyone

(5-SS)



To generate the LF representations we adjoin the NP to the topmost S node (creating a new S node) and leaving behind a co-indexed trace—in effect a bound variable.

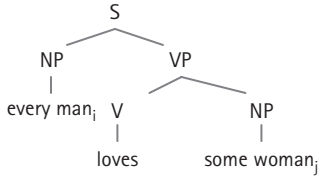
(5-LF)



Over the next decade or so a number of arguments were offered in support of QR and LF. Quite naturally, it was seen as a way of providing structural representations that could account for quantifier scope ambiguities. For example, consider sentence (6) and its SS representation.

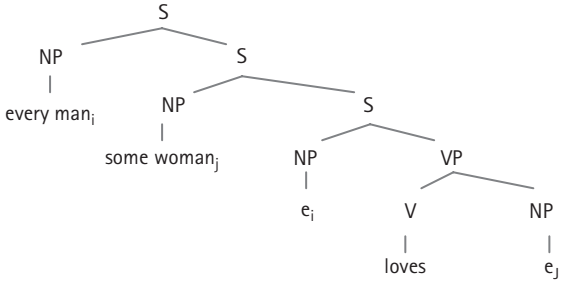
(6) Every man loves some woman

(6-SS)

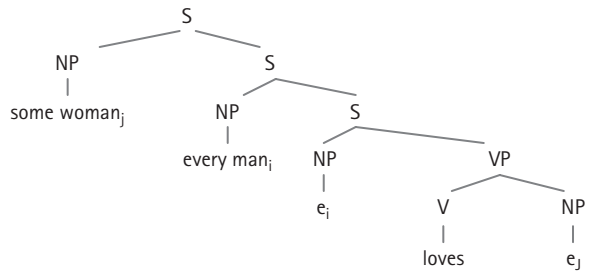


Given that either NP could raise first, this predicted two possible LF structures for the sentence, as indicated in LF1 and LF2.

(6-LF1)



(6 -LF2)



Another argument for QR driven by philosophical work in twentieth-century logic was that it could account for de re/de dicto ambiguity. To see this, consider (7) and the two resulting scope LF representations (7-LF1) and (7-LF2).[7] In the case of (7-LF1) the quantifier takes scope outside of the attitude verb generating the de re reading. In (7-LF2) it retains scope inside the attitude verb (by adjoining to the lower S) yielding the de dicto reading.

(7) John believes that a man is following him

(7-LF1) de re: [$_S$ (a man)$_i$ [$_S$ John believes that [$_S$ e_i is following him]]]
(7-LF2) de dicto: [$_S$ John believes that [$_S$ (a man)$_i$ [$_S$ e_i is following him]]]

The contributions from philosophers and philosophical methods were not always blatant importation of philosophical ideas. Many analyses were more subtle. For example,

Higginbotham (1980) argued that there was a contrast between (8) and (9), in that taking binding to work as indicated by the indices we judge (8) to be awful, but (9) to be significantly better (if you are having trouble getting the reading it helps to stress 'saw').

(8) *Who$_i$ did his$_i$ sister see e_i
(9) His$_i$ sister saw John$_i$

Higginbotham suggested that the relevant generalization here was the leftness condition discussed above.

Now, given this generalization, consider (10).

(10) His sister saw everyone

This sentence cannot mean that everyone is such that his sister saw him. The explanation is that if we propose QR, the leftness condition will automatically rule out the following

(10-LF) [(everyone)$_i$ [his$_i$ sister saw e_i]]

Facts like these are sometimes called "*weak* crossover" facts because they involve a case of LF movement in which the quantified expression crosses over the pronoun.

Another compelling argument for LF came from the phenomenon of "antecedent contained deletion." Consider (11).

[7] I've flattened the LF representations and removed some detail to economize on space.

(11) John suspected everyone that Mary did

If the deleted VP is simply reconstructed, we get the following.

(11-r) John suspected everyone that Mary *suspected everyone that Mary did*

But now have begun an infinite regress. An elided VP must be inserted for 'did' again (presumably the VP "suspected everyone that Mary suspected everyone that Mary did"), leaving yet another elided VP to be reconstructed.

Adapting a proposal from Sag (1976), May (1985) proposed the following. Suppose that QR takes place before the VP is reconstructed, so that the application of QR to (11) is (11-LF):

(11-LF) (everyone$_i$ that Mary did)$_i$ John suspected e$_i$

After reconstruction of the VP we now get (11-LFr) which has just the right interpretation (everyone who is such that Mary suspected them, is such that John suspected them):

(11-LFr) (everyone$_j$ that Mary suspected e$_j$)$_i$ John suspected e$_i$

Another interesting class of facts involves the "inverse linking" cases initially discussed in May (1977).

(12) Someone from every city despises it

May noted that this is ambiguous between a reading in which every city is such that someone from it despises it, and one in which a well-travelled person despises something. Notice that binding is not possible in the latter case, presumably because the quantifier does not c-command the pronoun.

(12-LF1) (every city)$_i$[[Someone from e$_i$] despises it$_i$]
(12-LF2) [[Someone from (every city)$_i$] despises it$_j$]

These examples are just a sample of cases in which philosophical concepts like quantifier scope and variable binding were incorporated into linguistic theory—not just at the periphery but in elements at the core of the theory and in ways that interacted in complex ways with the rest of the theory.

So far I've been talking about philosophical contributions that impinged upon syntax. The contributions to semantic theory have been greater by an order of magnitude. These contributions begin with the adoption of Frege's ideas about functions and Church's development of the lambda calculus, to the development of a semantics for modal logic by Kripke. It would be impossible to discuss all of these contributions, but I do want to call special attention to the impact of work on tense and the theory of events, because the philosophical contributions in these areas have extended well beyond generative linguistics and have had uptake even in descriptive linguistics.

For example, Reichenbach's (1947) work on tense logic (and the idea that a semantics of tense requires keeping track of a reference event R in addition to a speech event S and the event under discussion E) has been incorporated almost directly into descriptive linguistics. A good example of this would be in Comrie's (1985) book *Tense*.

No less important has been the use of event theory as initially developed by Davidson (1967b) and used in the development of the theory of thematic roles and the theory of the lexicon (see Grimshaw (1990), Hale and Keyser (1987, 1993), Pustejovsky (1995), Nirenberg and Raskin, (1987), Pustejovsky and Bergler (1991)), and in the theory of aspect (see Comrie 1976).

Finally, it would be a huge oversight to not remark on the contribution of philosophers to work in pragmatics. Pragmatics in some form or other has been part of linguistics forever, but the development of speech act theory by Austin and Searle and even more importantly the development of modern pragmatics by Grice (1989) has been monumentally important in linguistics, and is a part of virtually every introductory linguistics class taught today. While subsequent work (e.g. Sperber and Wilson (1986)) departs from Grice on important details it certainly remains theoretically grounded in Grice's work.

In sum, many traditional resources from twentieth-century philosophy of language found their way into linguistic theory, and while sometimes the mechanisms introduced were retasked and modified, they were nevertheless undeniably important. Some form of syntactic theory might have emerged without the contributions from philosophy, but it is really hard to imagine what syntactic theory would look like today without the contributions of philosophical tools and methods like quantifier scope, variable binding, etc. Furthermore, I don't think there is any question but that semantic theory as practiced in linguistics departments today would not be possible at all without the many philosophical resources it has adopted.

Not all of the work discussed above was ultimately successful or fruitful, but that is business as usual in a healthy science—there are false starts and promising if ultimately unsuccessful research projects. However on the whole, at least within generative linguistics, there is the general sense that the contributions from philosophy have been critical. There are exceptions.

Noam Chomsky has long been suspicious of attempts to introduce the philosophical notion of reference into linguistic theory (see the interview with Chomsky in Ludlow (2011a; appendix)) and he has likewise been cool towards related notions like external content. For all that, however, he has engaged philosophers on these issues and while Chomsky has rejected notions like reference (strictly speaking he considers the notion insufficiently sharpened to be deployed in a scientific theory of language), many other linguists have embraced the notion.

Interestingly, in rejecting referential semantics, Chomsky has perhaps unwittingly triggered a very interesting series of linguistic investigations. For example, Chomsky (1981) and Hornstein (1984) have argued that a referential semantics makes no sense from the perspective of linguistic theory, since linguistics doesn't distinguish beween a noun phrase like "the average family" and a noun phrase like "the coat in the closet". Yet, they say, no one thinks there are average families in the same sense there are coats in the closet. While I (2011a; chapter 6) was happy to bite this bullet, others have argued that we need to rethink our analysis of these constructions.

For example, Higginbotham (1985) argued that we should take "average" in "the average family" to be an adverbial element (like "quick" is in "quick cup of coffee"). Thus "The average family has 2.3 children" should be thought to have a syntactic form more along the lines of "on average, a family has 2.3 children". Alternative proposals have been offered

by Carlson and Pelletier (2002) and Stanley and Kennedy (2009). This is not the place to explore the various proposals on the table; my only point here is that even where philosophical proposals have been rejected they have led to some interesting and productive proposals with tangible empirical consequences.

I think it is fair to say that whatever controversy there may be about the contribution of philosophical methodology within linguistics, the controversies arise on the level of individual proposals and are not global worries on the level of "who invited the philosophers."

As we will see in the next section, however, there *are* global worries about the introduction of linguistic methodology into philosophy. As we will also see, I consider those worries to be misplaced.

3. Linguistic Methodology in Philosophy

Just as there has been significant application of traditional philosophical tools in linguistics, the borrowing works in the other direction as well. In some cases we get a nice feedback loop, with philosophical tools like quantifier scope being imported into linguistics, and then linguistic work on scope and binding being imported back into philosophy. Let's look at some concrete examples.

3.1 Indefinite Descriptions

Russell's theory of descriptions was a key tool in analytic philosophy over the past 100 years, since it could be wielded both as a tool to minimize commitment to suspect entities (round squares and unicorns) and as part of a story about our epistemic access to the world (some thing are known by acquaintance and others by description). Russell's story came under pressure from Donnellan (1966) who observed that there are many cases where we use descriptions referentially, and other cases where we use them in a way that Donnellan called "attributive." Kripke (1977) and Neale (1990) pushed back on this thesis, arguing that the referential uses of description were not a feature of the semantics of descriptions but rather the way that descriptions were used.

This line of argument in turn led to a number of very interesting arguments from linguists (e.g. Fodor and Sag (1982)), who tried to make the case that there is good linguistic evidence for the referential/attributive distinction being the reflex of semantic phenomena. One piece of evidence for this was the idea that certain facts about sentences with indefinites could not be accounted for by allowing indefinites to take wide scope, since this would require that indefinites violate "scope islands"—they would have to move in a way that syntax simply doesn't allow. Ludlow and Neale (1991) and King (1988) argued against this, utilizing standard linguistic counterexamples, as well as linguistic probes into effects that might obscure linguistic data (see Ludlow (2011a; Chapter 3) for discussion). In other words, it became necessary to engage in empirical linguistic research to determine whether a particular analysis of descriptions was plausible. This is not an isolated instance;

for all practical purposes the philosophical and linguistics literature on descriptions has become completely intertwined.[8]

3.2 Complex Demonstratives

In section 2 of the chapter I showed how the contribution of philosophers in linguistic theory (in particular the contribution of ideas relating to scope and binding) had provided resources that led to constraints like the "leftness" condition and explanations for antecedent contained deletion. Interestingly King (2001) in turn borrowed these new linguistic discoveries back for his work in philosophy and showed how they had philosophical consequences for the theory of complex demonstratives. In particular, King wanted to show that certain kinds of complex demonstratives were quantificational and not referential. He gave several arguments for this conclusion, but two of the arguments leaned on the notions of weak crossover and antecedent contained deletion that we discussed in section 2.

For example, consider (13).

(13) His mother loves that man in the goatee

King believed that "his" and the complex demonstrative cannot have the same reference (unless the mother fails to realize it is her son), and this is predicted from the LF we would get if the complex demonstrative is a quantified expression:

(13-LF) (that man in the goatee$_i$) [His$_i$ mother loves e$_i$]

As before this would be ruled out by the leftness condition. King likewise drew on antecedent contained deletion to explain why complex demonstratives work in constructions like (14).

(14) Tiger birdied that hole that Michael did

The LF would be as follows:

(14-LF) (that hole that Michael did)$_i$ [Tiger birdied e$_i$]

And when the VP is reconstructed we end up with the following.

(14-LFr) (that hole that Michael birdied e$_j$)$_i$ [Tiger birdied e$_i$]

My concern here isn't with whether King got the analysis right; it is rather a point about the relation between linguistic and philosophical methodology—at least in the domain of the philosophy of language. When philosophical resources were introduced into linguistic theory one might not have expected the resulting linguistic developments (the discoveries and explanations for weak crossover and antecedent contained deletion) to play a role

[8] See my (2011b) article "Descriptions" in *The Stanford Encyclopedia of Philosophy* for a detailed discussion of this.

in philosophy, yet clearly they have (King's use of these resources is not the only instance). This is another example where we get a two-way flow between linguistics and philosophy.

It would be a mistake to suppose that this two-way flow can only be found in the philosophy of language. I also believe that we get an important two-way flow between linguistics (in particular semantics) and the metaphysics of time.

3.3 Tense and the Philosophy of Time

I noted in section 2 of this chapter that linguists had adopted an account of tense from Reichenbach (1947). I might have noted that for the most part they have also rejected accounts of tense like Prior's (1967, 1968) in which tense is treated as an operator.

One key reason that linguists prefer the Reichenbachian story is that it provides them a handy way to account for temporal anaphora. Temporal anaphora is like pronominal anaphora in that there is a linking to a referent that occurs earlier in the discourse, but in the case of temporal anaphora it is a temporal reference point. Partee (1984) provides the *locus classicus* of these cases, including example (15).

(15) I turned off the stove

(15) is typically not uttered to say that the speaker turned off the stove once in his life, but rather that the stove was turned off during some relevant time or during some relevant interval (say when the speaker finished preparing dinner today). Linguists have also been happy to endorse the apparent metaphysical consequences of this—a commitment to a form of four-dimensionalism (if I am making reference to past events then they must exist and presentism must be false). In Ludlow (1999, forthcoming) I tried to make the case for an account of tense more in the spirit of Prior, and I argued that tense is an ineliminable property for the semantics of natural language, and by extension for our metaphysics of time.

The argument was that that although such properties may not be of use in the statement of a physical theory, they are still ineliminable in our current explanations of human action. The argument follows familiar ground.

Suppose I am sitting in my office, aware that I have a meeting with my boss at 4:00 o'clock. I may have written in my calendar that I have the meeting at 4:00. Now let's suppose that is 4:00 but I don't realize it is, so I do not get up and go to the meeting. But then I look at a clock and see the time. I realize that my meeting is happening at that moment, so I get up and run to the meeting.

Now consider two possible four o'clock utterances:

(16) I have a meeting with my boss at four o'clock.
(17) Oh no, I have a meeting with my boss now!

In the scenario we imagine, the belief I have when I utter (16) is not sufficient to get me up and off to the meeting, but the belief I have when I utter (17) *is* sufficient to get me up and off to the meeting. Obviously there is something different going on when I make those two utterances, and you might believe that the difference is reflected in the semantics of

those two utterances—that the difference is reflected in the semantic content of my utterances. And if semantics really is about language/world connections, then there are distinct properties being expressed in those two utterances. In the second case I am expressing a perspectival property.

So far I've suggested that metaphysical tense is necessary to account for the semantics of the temporal indexical "now," but the same point really extends to any kind of tensed expression—for example, one in which we situate an event in the past.

Consider A. N. Prior's famous (1959) example of (18) contrasted with (19). Let's suppose both are uttered at four o'clock GMT, on 11-11-11.

(18) I am thankful that my root canal is over with
(19) I am thankful that my root canal is earlier than four o'clock GMT, on 11-11-11.

Prior held that it isn't sufficient to say that someone who utters (18) is grateful the root canal was complete at four o'clock GMT, on 11-11-11—that just isn't enough to explain the relief (thankfulness) on the part of the speaker. The speaker is grateful that the root canal is safely in her past—the root canal is temporally removed from her egocentric position—and is not thankful that the root canal is later than some particular time point. The example also shows that it is not just accounts of human action that requires these perspectival properties but that accounts of human emotions may require them as well. If this is reflected in the semantics, and semantics expresses language/world relations, then once again we are forced to posit and rely on perspectival properties.

This is certainly not the only example of an attempt to argue from the semantics of tense to the metaphysics of time, but these examples are sufficient to allow us to raise the key question: Is this be the right way to do metaphysics? Ted Sider, drawing attention to another attempt to draw metaphysical conclusions from temporal language, has argued that it is not:

> [F]our-dimensionalism is a *metaphysical* thesis about the nature of persisting objects. It is *not* a thesis about language, nor about the analysis of predicates of continuants, nor about the conceptual epistemic priority of predicates of states and predicates of continuants…
>
> The difference between thing-talk and process-talk in no way undermines four-dimensionalism. It is consistent with things and events being in the same ontological category that natural language contains different ways of speaking of things and events. Natural language contains different vocabulary for speaking of persons and inanimate physical objects, but this is no argument against materialism. Nor is the oddness of saying that my thought is spatially located in my brain a compelling argument against the mind–brain identity theory. The objection might have bite if four-dimensionalism were a thesis of ordinary language philosophy, but it seems ineffective against the metaphysical thesis that I uphold.
>
> (Sider 2001; 211–12)

I believe that Sider's concerns reflect a misunderstanding about the nature of semantics as well as a misunderstanding about the nature of the project I at least am engaged in. The idea is not to read metaphysics or anything else off of linguistic *forms*. When we are engaged in the semantics of natural language we are already engaged in language/world relations. Semantics would simply be impossible if we did not have some grasp on the structure of the world.

Accordingly, when Sider says that "It is consistent with things and events being in the same ontological category that natural language contains different ways of speaking of things and events," he is missing the point. The semantics of natural language doesn't care that we use a feminine form when we speak of some individuals and a masculine form (or neutral form) for others. That is a distinction that doesn't have semantic bite. Sider gives the example of the difference between thing-talk and process-talk, and argues that "the difference between thing-talk and process-talk in no way undermines four-dimensionalism." Whether Sider is right about this depends entirely on the nature of thing-talk and process-talk and what the semantics of those constructions must look like. Semantics might completely ignore the difference. On the other hand, if it turns out that so-called process-talk involves some irreducible progressive aspect, then perhaps we should not be so eager to dismiss the idea that the talk is the reflex of something metaphysically important.

For example, suppose that there were elements of process talk that we could not reduce to thing-talk and still make sense of our attitude attributions or our explanations of human action. That is, suppose that if we give a regimented semantics for process-talk we can't explain how the contents of our utterances hook up with the theory of human action. Should this not give us pause?

Let's go back to the example of tense. I made the case that there is such a thing as metaphysical tense and that metaphysical tense can be understood in terms of first person perspectival properties. However, I did not argue that we can make a case for tense by arguing from the existence of some syntactic notion of tense. To the contrary, very few languages in the world have grammatical tense, and the various languages of the world have many different strategies for expressing tense. Some languages use modals (like we do in English to express the future), some languages use evidential markers, others (like Chinese and the Slavic languages) use aspect.

So why did I think that tense was ineliminable? The case for metaphysical tense is not made by projecting some abstract feature of the syntax of language onto the world; to the contrary metaphysical tense comes first on this story and different languages have different strategies for talking about metaphysical tense. The case for ineliminability comes from the observation that once we squeeze the tense out of our semantics for these utterances we can no longer account for our attitude attributions or our explanations of human behavior.

Physics, of course, may not care about attitude attributions or explaining human behavior, and hence it may have no need for the kinds of perspectival properties that other sciences need. That is fine. The mistake comes when we suppose that all metaphysically kosher properties must be of utility to the current state of physics. The semantics of natural language cannot afford to be so restrictive.[9]

We can put the point like this: Just as linguistics and philosophy co-evolved, so too our natural language semantics and human explanatory enterprises co-evolved. We needn't get into a dispute about which came first. You can say that semantics is the way it is because the world is the way that it is, or you can say that semantics constrains the way in which

[9] Am I supposing that the semantics of natural language and the theory of human action are "special sciences" in the sense of Fodor (1974)? I am saying that other sciences need properties for which we do not yet understand how they can or will supervene on the basic properties of physical theory. For current purposes I'm neutral on whether these sciences must ultimately be unified.

we conduct our inquiries into the way the world hangs together. I don't care: either way the point is that the semantics of natural language incorporates our theory of everything (because it has to talk about everything to an arbitrary level of detail). This means that it is very dangerous business to start denying semantics important resources because they are not needed in, for example, contemporary physics. Semantics has to worry about a lot more than the language of physics.

If semantics incorporates our theory of everything one might ask why we bother with the linguistic formulations at all: why not cut directly to the world itself? To put it another way, why talk about language when we can talk about the world instead and not lose anything?

The reason we attend to linguistic formulations and pay attention to the semantics is simply because it affords us a level of precision in our discussion of these matters. It prevents people from surreptitiously utilizing semantic properties (e.g. metaphysical tense) while at the same time denying the need for such properties. For example, it is easy enough to wave a hand and say that the use of a progressive or tense has no reflex in reality, but the semanticist is there to say, "Wait a minute, slow down, you are relying on those properties whether you admit it or not."

This general point holds in other applications of semantics to core philosophical issues. Consider the case of contextualism in epistemology.

3.4 Contextualism in Epistemology

One of the dominant theories proposed in epistemology over the last few decades has been contextualism—a view that holds that standards of knowledge can vary from context to context (see, for example, DeRose (1992, 1995), and Cohen (1999)). Typically, contextualism is grounded in an observation about our knowledge-reporting practices, and sometimes it is taken to be a thesis about the lexical semantics of the verb "knows"—does the verb have a hidden argument position for standards of knowledge?

Everyone agrees that one can modify knowledge reports, saying "I know by the standards of x", but the linguistic question is whether we are simply modifying the knowledge report with an adverb-like modifier or whether this is a legitimate argument position, so that "knows" is at a minimum a three-place verb that takes an experiencer (knower), a content (what is known), and a standard (for example by the standards of an epistemology class or the standards of a courtroom).

In Ludlow (2005) I argued that you could make a linguistic case for such argument positions, in part because of the non-iterability of these standards—that is, just as it is odd to say "John ate dinner with a knife with a fork" (it seems you only get one instrument unless you insert a conjunction), it is odd to say "John knows he has feet by the standards of law by the standards of an epistemology class on skepticism". Meanwhile Stanley (2005) offered linguistic arguments going in the other direction. For example, he observed that if there really is an argument position for standards, you would expect binding facts to appear. So, for example, contrast the following.

(20) Everyone went to a local bar
(21) Everyone knows that Chesner has feet.

(20) allows an interpretation in which each person went to a bar that is local to that person, but (21) does not seem to allow the standard of knowledge to shift with respect to person.[10]

My point here is not to hash out the debate but simply to give a sample of the kinds of semantic arguments that have been offered on each side of the debate. The question again is this: Is this any way to do epistemology?

I certainly think that it is. One can of course modulate the meaning of "knowledge" any way one pleases and presumably one can do this in a way that only an Unger-style conception of "knows" is expressed—that is, a conception in which the only admissible standard for "knows" would be something like Cartesian certainty, where that standard is baked into the very meaning of "knows". You might even argue that this modulated meaning plays an important role in theorizing. But again the semanticist has more to worry about than the theoretical interests of established epistemologists. The semanticist is responsible for making sense of all epistemic language, and this includes our uses of verbs like "know", nouns like "justification", adverbs like "probably" (in the epistemic sense), and adjectives like "certain".

Of course, it is entirely possible to argue that the bulk of our uses of epistemic language are only so much loose talk—that what we are saying is strictly speaking false but that we are pragmatically licensed to say such things. Whether or not we find such claims persuasive (I personally do not) notice that they are in themselves claims that fall under the purview of the semantics of natural language: they are claims about the semantics/pragmatics divide. Such claims are in the domain of semantics if anything is.

4. Two Methodologies or One?

As we have seen, some of the key concepts in philosophy and linguistics have migrated back and forth between these enterprises. One of the examples we have looked at—the notion of variable binding—seemed to start in philosophy, then get incorporated into linguistics, and then was ultimately imported back into philosophy in its linguistic dressing (e.g. in Stanley (2005) and King (2001)).

But given this migratory methodology, is there really any point in characterizing the talk of (for example) binding relations as being inherently philosophical or inherently linguistic? I think not. I believe the distinction is artificial. The temptation is to think that the philosophical component must have come first because it was developed BC (before Chomsky) but this overlooks the way that the study of language and philosophy were intertwined through the medieval period and on into the nineteenth century. Even the notion of quantifier scope, which we might take to have been philosophical in origin is wrapped

[10] I'm not trying to settle the issue here and am just giving an example of appeals to semantics in philosophy, but for the record I don't think this argument works. The contextualist would have the standards fixed by the context of assessment (the time and place in which the knowledge report is uttered), and even if we are quantifying over many individuals in (21) there is still only one context of assessment. Stanley's argument stands in need of repair on this point.

up in medieval theories of grammar and logic, and these enterprises in large measure, coevolved.[11]

More controversially, I don't think there is a clear sense in which we can say that a particular bit of reasoning about the existence of tensed properties is linguistic or whether it is metaphysical—even in the philosophy of time. I think "coevolved" is an apt way to describe the relation between linguistic and philosophical methodology. But if this the case then what really is the difference between the methodologies? No doubt the difference is largely institutional at this point. Work that takes place in a linguistics department typically gets classified as linguistics and work that takes place by someone in a philosophy department typically gets classified as philosophy.

Obviously there is good work in philosophy that makes no explicit reference to semantics, but I think that this is largely because the role of semantics in the reasoning is enthymematic; the semantic argumentation is part of the background and often not explicitly proffered. Of course this kind of enthymematic reasoning can get one in trouble—not being explicit about the steps taken can be a recipe for error.

I've argued that semantics includes as part of its subject matter our most complete theory of the world. It should therefore come as no surprise that that methodology in linguistics and in particular the semantics of natural language is intimately intertwined with our methodology in numerous areas of philosophy. This is reflected, not just in contemporary semantics and philosophy, but also in the very history of both these areas of inquiry. If I'm right it also suggests that attention to detail in the semantics of natural language can save us from error and oversight in areas ranging from the philosophy of language to metaphysics and epistemology. Indeed, it is hard to see how progress *could* be made in philosophy without attention to semantic details.

References

Buridan, J., 1966. *Sophisms on Meaning and Truth*. Trans. T. K. Scott. New York: Appleton-Century-Crofts.

Carlson, G., and F. J. Pelletier, 2002. "The Average American has 2.3 children." *Journal of Semantics* 19, 73–104.

Chomsky, N., 1973. "Conditions on Transformation." In S. Anderson and P. Kiparsky (eds.) *A Festschrift for Morris Halle*. New York: Holt, Reinhart, and Winston, 232–86. Reprinted in Chomsky (1977).

Chomsky, N., 1975. *The Logical Structure of Linguistic Theory*. Chicago: University of Chicago Press. (Originally appeared in unpublished manuscript form in 1955.)

Chomsky, N., 1976. "Conditions on Rules of Grammar." *Linguistic Analysis* 2, 303–51. Reprinted in Chomsky (1977).

Chomsky, N., 1977. *Essays on Form and Interpretation*. Amsterdam: Elsevier NorthHolland.

Chomsky, N., 1981. *Lectures on Government and Binding*. Dordrecht: Foris Publications.

Cohen, S., 1999. "Contextualism, Skepticism, and the Structure of Reasons." *Philosophical Perspectives 12: Epistemology*, 57–89.

Comrie, B., 1976. *Aspect*. Cambridge: Cambridge University Press.

[11] See, for example, some of the issues discussed in Buridan (1966).

Comrie, B., 1985. *Tense*. Cambridge: Cambridge University Press.
Davidson, D., 1967a. "Truth and Meaning." *Synthese* 17, 304–23.
Davidson, D., 1967b. "The Logical Form of Action Sentences." In N. Rescher (ed.) *The Logic of Decision and Action*. Pittsburgh: University of Pittsburgh Press, 81–95.
den Dikken, M., R. Larson, and P. Ludlow, 1996. "Intensional Transitive Verbs." *Rivista di Linguistica* 8, 331–48. (Abridged version reprinted in P. Ludlow (ed.) *Readings in the Philosophy of Language*, Cambridge: MIT Press, 1997).
DeRose, K, 1992. "Contextualism and Knowledge Attributions." *Philosophy and Phenomenology Research* 52, 913–29.
DeRose, K., 1995. "Solving the Skeptical Problem." *Philosophical Review* 104, 1–52.
Donnellan, K. S., 1966. "Reference and Definite Descriptions." *Philosophical Review* 77, 281–304.
Dowty, D., R. Wall, and S. Peters, 1981. *Introduction to Montague Semantics*. Dordrecht: D. Reidel.
Dyson, F., 2012. "What Can You Really Know?" *The New York Review of Books*, Nov. 8, 2012. (<http://www.nybooks.com/articles/archives/2012/nov/08/what-can-you-really-know/>). Accessed September 24, 2015.
Fiengo, R., 1977. "On Trace Theory." *Linguistic Inquiry* 8, 35–61.
Fodor, J., 1970. "Three Reasons for Not Deriving 'Kill' from 'Cause to Die'." *Linguistic Inquiry* 1, 429–38.
Fodor, J., 1974. "Special Sciences (or: The Disunity of Science as a Working Hypothesis)." *Synthese* 28, 97–115.
Fodor, J. D., and I. Sag, 1982. "Referential and Quantificational Indefinites." *Linguistics and Philosophy* 5, 355–98.
Grice, P., 1989. "Logic and Conversation." In Grice (ed.) *Studies in The Way of Words*. Cambridge: Harvard University Press, 22–40.
Grimshaw, J., 1990. *Argument Structure*. Cambridge: MIT Press.
Hale, K., and Keyser, J., 1987. "A View from the Middle." Lexicon Project Working Papers 10, Center for Cognitive Science, MIT.
Hale, K., and Keyser, J., 1993. "Argument Structure." In K. Hale and J. Keyser (eds.) *The View from Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*. Cambridge: MIT Press, 53–109.
Heim, I., and A. Kratzer, 1998. *Semantics in Generative Grammar*. Oxford: Blackwell.
Higginbotham, J., 1980. "Pronouns and Bound Variables." *Linguistic Inquiry* 11, 679–708.
Higginbotham, J., 1985. "On Semantics." *Linguistic Inquiry* 16, 547–94.
Hornstein, N., 1984. *Logic as Grammar*. Cambridge, MA: MIT Press.
Katz, J., and J. Fodor, 1963. "The Structure of a Semantic Theory." *Language* 39, 170–210.
Katz, J., and P. Postal, 1964. *An Integrated Theory of Linguistic Description*. Cambridge: MIT Press.
King, J., 1988. "Are Indefinite Descriptions Ambiguous?" *Philosophical Studies* 53, 417–40.
King, Jeffrey 2001. *Complex Demonstratives: A Quantificational Account*. Cambridge: MIT Press.
I. Heim, and A. Krazter, 1998. *Semantics in Generative Grammar*. Oxford: Blackwell Publishing.
Kripke, S., 1977. "Speaker's Reference and Semantic Reference." In P. French, T. Uehling, and H. Wettstein (eds) *Contemporary Perspectives in the Philosophy of Language*. Minneapolis: University of Minnesota Press, 255–76.

Larson, R., M. den Dikken, and P. Ludlow, 1997. "Intensional Transitive Verbs and Abstract Clausal Complementation." Manuscript, SUNY Stony Brook, and Frei Universiteit Amsterdam. Available online at <http://semlab5.sbs.sunysb.edu/~rlarson/itv.pdf>. Accessed October 19, 2015.

Larson, Richard, and Gabriel Segal, 1995. *Knowledge of Meaning: Semantic Value and Logical Form*. Cambridge: MIT Press.

Lewis, D., 1972. "General Semantics." In D. Davidson and G. Harman (eds.) *Semantics of Natural Language*, Dordrecht: D. Reidel, 169–218.

Ludlow, P., 1999. *Semantics, Tense, and Time: An Essay in the Metaphysics of Natural Language*. Cambridge: MIT Press.

Ludlow, P., 2005. "Contextualism and the New Linguistic Turn In Epistemology." In G. Preyer and G. Peter (eds.) *Contextualism in Philosophy*. Oxford: Oxford University Press, 11–50.

Ludlow, P., 2011a. *The Philosophy of Generative Linguistics*. Oxford: Oxford University Press.

Ludlow, P., 2011b. "Descriptions." In Edward N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy (Winter 2011 Edition)*. URL = <http://plato.stanford.edu/archives/win2011/entries/descriptions/>. Accessed September 24, 2015.

Ludlow, P., forthcoming. "Tensism." In L. de Saussure & K Jaszczolt (eds), *Time: Language, Cognition and Reality*. Oxford: Oxford University Press.

Ludlow, P., and S. Neale, (1991). "Indefinite Descriptions: In Defense of Russell." *Linguistics and Philosophy* 14, 171–202.

May, R., 1977. *The Grammar of Quantification*. unpublished Ph.D. dissertation, MIT.

May, R., 1985. *Logical Form: Its Structure and Derivation*. Cambridge: MIT Press.

Montague, R.1974. *Formal Philosophy: Selected Papers of Richard Thomson*. (Richmond Thomason, ed.). New Haven: Yale University Press.

Neale, S., 1990. *Descriptions*. Cambridge: MIT Press.

Newmeyer, Frederick, 1986. *Linguistic Theory in America*, 2nd edition. San Diego: Academic Press.

Nirenberg, S., and V. Raskin, 1987. "The Subworld Concept Lexicon and the Lexicon Management System." *Computational Linguistics* 13, 276–89.

Partee, B., 1984. "Nominal and Temporal Anaphora." *Linguistics and Philosophy* 7, 243–86.

Postal, P., 1971. *Cross-Over Phenomena*. New York: Holt, Reinhart, and Winston.

Prior, A.N., 1959. "Thank Goodness That's Over." *Philosophy* 34, 12–17.

Prior, A. N., 1967. *Past, Present and Future*. Oxford: Oxford University Press.

Prior, A. N., 1968. *Time and Tense*. Oxford: Oxford University Press.

Pustejovsky, J., 1995. *The Generative Lexicon*. Cambridge: MIT Press.

Pustejovsky, J., and S. Bergler (eds.), 1991. *Lexical Semantics and Knowledge Representation*. Berlin: Springer-Verlag.

Reichenbach, H., 1947. *Elements of Symbolic Logic*. New York: Macmillan.

Sag, I., 1976. "A Note on Verb Phrase Deletion." *Linguistic Inquiry* 7, 664–71.

Sider, T., 2001. *Four-Dimensionalism: An Ontology of Persistence and Time*. Oxford: Oxford University Press.

Sperber, D., and D. Wilson, 1986. *Relevance: Communication and Cognition*. Oxford: Blackwell.

Stanley, J., 2005. *Knowledge and Practical Interests*. Oxford: Oxford University Press.

Stanley, J., and C. Kennedy, 2009. "On 'Average'." *Mind* 118, 583–646.

Wasow, T., 1972. *Anaphoric Relations in English*. PhD Dissertation, Linguistics, MIT.

CHAPTER 26

HISTORY OF IDEAS
A Defense

FREDERICK C. BEISER

1. Philosophy and the History of Ideas

The history of ideas, as Arthur Lovejoy defined it in 1940 in the opening number of the *Journal of the History of Ideas,* is "a duly analytical and critical inquiry into the nature, genesis, development, diffusion, interplay and effects of the ideas which generations of men have cherished, quarrelled over, and apparently been moved by."[1] The ideas that were the object of Lovejoy's study he understood in a very broad sense. They could be ideas from any discipline or activity, the ideas of science, art, literature, religion, and politics, and last, but certainly not least, philosophy. Lovejoy's definition, coming from a great master of the discipline and in the inaugural issue of its signature journal, has some authority of its own; it is as good a definition as we are likely to get.

Though Lovejoy did not live to see it, the history of ideas gradually split into two branches: intellectual history, which studies the ideas of a particular person, movement, school, or epoch; and the history of concepts, or what the Germans call *Begriffsgeschichte*, which studies the development and variations of a concept in different persons, movements, schools, or epochs. *Begriffsgeschichte* attempts to do for concepts—whatever they are—what etymology does for words.[2]

The history of ideas, in both its branches, has its origins in nineteenth-century German historicism.[3] It grew out of the attempt to apply the methods of history, as developed and

[1] Arthur Lovejoy, 'Reflections on the History of Ideas', *The Journal of the History of Ideas,* 1 (1940), 3–23, p. 8.

[2] See the journal *Archiv für Begriffsgeschichte* (Bonn: Bouvier, 1955f), begun by Erich Rothacker and now edited by the Akademie der Wissenschaften. On the origins of "*Begriffsgeschichte",* see the article by H. G. Meier in *Historisches Wörterbuch der Philosophie*, ed. Joachim Ritter et al. (Basel: Schwabe & Co. 1971), I, 788–808. The entire dictionary is a paradigmatic product of the discipline.

[3] On this tradition, see my *The German Historicist Tradition* (Oxford: Oxford University Press, 2011).

refined by such nineteenth-century historians as Barthold Niebuhr (1776–1831), Leopold Ranke (1795–1886), Johann Droysen (1838–1908) and Wilhelm Dilthey (1883–1911), to intellectual life. With the same methods that we use to produce a history of Rome or Prussia, these thinkers believed, we can also have a history of philosophy, science, or art. These methods were *critical* (viz., assessing the authenticity of sources), *genetic* (viz., determining origins and causes), *holistic* (viz., placing historical phenomena in context and crossing artificial disciplinary boundaries), and *individualizing* (viz., determining what was unique to a historical phenomena). Later historians of ideas in the twentieth century, not least the contributors to Lovejoy's journal, very much followed these methods, even if they were unaware of their source.

There is a tendency among Marxists and positivist historians to dismiss the history of ideas as a bogus discipline because it is based on hypostasis. Ideas are not, we are told, things, and so they have no histories.[4] But this objection is bunkum. It confuses historicism with the idealist theory of history, especially with Hegel's and Humboldt's metaphysics, which makes the idea the ultimate force behind history. Thinkers in the historicist tradition distanced themselves from metaphysics, especially Hegel's philosophy of history: nominalists by inclination, they too rejected ascribing causal agency or essential meanings to ideas. Their intention was to make intellectual history as rigorous an empirical discipline as history in general.

The history of philosophy is a branch of the history of ideas, namely, the history of *philosophical* ideas. Some object to placing the history of philosophy within the history of ideas because it will deprive it of its critical or philosophical dimension. But it is worth noting that the great practitioners of the history of ideas never saw it as an uncritical enterprise. Lovejoy pleaded that intellectual historians should never forget that their texts contain claims to truth which are honored only through criticism.[5]

Still, the history of philosophy is a *peculiar* branch of the history of ideas. Other branches allow for a critical component; they do not demand it. The art historian is not required to be a good critic; the religious historian need not be a seer; and the historian of science is not expected to be a good scientist. The historian of philosophy, however, is supposed to be a good philosopher. The evaluative or critical dimension is not only permitted but required. The philosopher is expected, not only to understand texts as historical documents, but also to criticize them as philosophical doctrines. Why this extra demand is placed on the philosophical historian is not so easy to understand. Philosophy, it is often said, has an especially intimate connection with its past. A contemporary philosopher develops his or her ideas by engaging with past philosophers in ways in which contemporary scientists, artists, or critics do not. Just as the historian of philosophy is expected to know philosophy, the philosopher is supposed to know the history of his subject. While an economist or physicist is

[4] The latest purveyer of this objection is Skinner, who refuses to accept that an idea can be "an appropriate unit of historical investigation". See his "Meaning and Understanding in the History of Ideas", *History and Theory* 8 (1969), 3–53, p. 31. Skinner maintains there is no "determinate idea" but only "a variety of statements made with words by a variety of agents with a variety of intentions", so that "there is no history of the idea to be written" (p. 36). But most advocates of *Begriffsgeschichte* would not disagree with Skinner that there is no such thing as the "essential meaning" of an idea. Their aim is to trace change in meaning through change in use.

[5] See Lovejoy, "Reflections", pp. 19–20.

not expected to read Smith or Newton, a philosopher should read Plato or Aristotle. This is perhaps because philosophical ideas are not so prone to obsolescence, refutation, or falling out of fashion as artistic, scientific, or literary ones.

Whatever the reason for the peculiarity of the discipline, it imposes special burdens upon its practitioners. It is not so easy to be both philosopher and historian. History and philosophy are very different disciplines, which require very different skills to solve very different kinds of problems. Asking for a good historian of philosophy is like asking for someone who is both a good psychoanalyst and plumber. The philosopher is concerned with *Begriffsspiel*, with the validity of arguments, with the logical structure of ideas (their premises and conclusions, their relations of compatibility and incompatibility with other ideas), and he does not have to worry about their origins or causes. The historian, however, is concerned precisely with the origins and causes of ideas and arguments, and it does not matter to him whether the ideas are true or false, or whether the arguments are valid or invalid. To use an old Kantian distinction: the philosopher deals with the *quid juris?*—"What right or justification do I have for a belief?"—while the historian answers the *quid facti?*—"What are the causes for such a belief?"

The need to juggle these conflicting demands, to answer these very different questions, is a perennial challenge for the historian of philosophy. While historians complain that his work is too philosophical, philosophers gripe that it is too historical. There are indeed serious dangers in leaning too far in one direction or the other. If a historian of philosophy is too historical, he flirts with *antiquarianism,* i.e., talking about the dead for their own sake when that interests no one; and if he is too philosophical, he runs the risk of *anachronism*, i.e., attributing to the past the interests of the present.

Given the very different concerns of philosophy and history, it should not be surprising that the history of philosophy has a tendency to fracture into historical and philosophical halves. This is exactly what happened to the discipline in Britain in the second half of the twentieth century. Some made it almost entirely philosophical, others almost completely historical. We are still living with this split today; everyone has felt it, though few are aware of its origins and context. So it is now time that I tell a story.

2. A Disasterous Divorce

Once upon a time, more than a half century ago, the young Peter Strawson gave his lectures on Kant's *Critique of Pure Reason* at Oxford. These lectures, begun in the late 1950s, came to fruition in 1966 when Strawson first published his *The Bounds of Sense*.[6] Strawson wanted to read Kant's book on purely philosophical terms, to examine and appraise its main arguments by engaging directly with the text in the Kemp Smith translation; he made no attempt to read it in German, to consult the commentaries on it, or to study the many controversies about it, in the nearly two centuries since its publication.[7] To many at the

[6] P. F. Strawson, *The Bounds of Sense: An Essay on Kant's Critique of Pure Reason*. (London: Metheun, 1966).

[7] See the "Preface" to *The Bounds of Sense*, p. 11.

time, Strawson's work seemed liberating: it was as if one could get to the philosophy pure and simple without having to take on the onerous tasks of learning German or a tradition of scholarship.

The precedent set by Strawson was followed by Jonathan Bennett, who published his *Kant's Analytic* also in 1966 and then his *Kant's Dialectic* in 1974.[8] Bennett shared in the excitement of those years, attempting to tackle Kant's text directly without having to take on the burden of historical scholarship. Brusquely, he pushed aside the opinion of one reviewer who wrote that he was "more interested in what Kant ought to have thought than in what he actually did think."[9] By the early 1970s Bennett and Strawson had become the new orthodoxy in Oxford, where their books became "recommended" (i.e. required) reading for the "Special Paper" on Kant's *Critique of Pure Reason*.[10]

Around the same time as Strawson and Bennett were liberating the history of philosophy from history, some young historians in Cambridge were liberating it from philosophy. In the late 1960s, Quentin Skinner advocated an approach to the history of ideas that stressed the importance of historical understanding over philosophical criticism.[11] His approach was later refined and radicalized by James Tully in the 1980s.[12] Their chief goals were to recover the original intentions of the author, to place him in historical context, and to understand him in his own terms. Because of their concern to recover intention and context, Skinner and Tully were skeptical of the value of philosophical criticism, which, they feared, would be only anachronistic.

Remarkably, these very different approaches created little friction. There was no debate about methodological issues, though there well should have been. But, then again, Oxford and Cambridge have never had much to do with one another, and it would make little difference if they were on different continents rather than just sixty miles apart. It is tempting to say with Richard Rorty that there is really no conflict between these methods, that they simply have different purposes.[13] But this irenic strategy sweeps too much under the carpet. The Cambridge historians, if they only had made a fleeting glance toward their colleagues in philosophy, would have questioned the value of not only their philosophical criticism, but also their a-historical interpretations of texts.

[8] Jonathan Bennett, *Kant's Analytic* (Cambridge: Cambridge University Press, 1966) and *Kant's Dialectic* (Cambridge: Cambridge University Press, 1974).

[9] Bennett, *Kant's Dialectic*, p. viii.

[10] In the interests of full disclosure, the present author should reveal that he was one undergraduate subjected to this regimen, and that his rebellion against it set the course of his career. The proscribed texts, "read in the loo", that fuelled this rebellion were some classics of the German historicist tradition, especially Haym and Dilthey, which gave him a very different idea of how to pursue the history of philosophy.

[11] See Quentin Skinner, "Meaning and Understanding", 3–53. See also his "Social Meaning and the Explanation of Social Action", in *Philosophy, Politics and Society* (fourth series), eds. Peter Laslett, W. G. Runcimann, and Quentin Skinner (Oxford: Blackwell, 1972), 136–57; and "Motives, Intentions and the Interpretation of Texts" in *On Literary Intention*, ed. D. Newton de Molina (Edinburgh: Edinburgh University Press, 1976), 210–21.

[12] See James Tully, "The pen is a mighty sword', in *Meaning and Context,* ed. James Tully (Cambridge: Polity Press, 1988), 7–25.

[13] This is the position of Richard Rorty, "The historiography of philosophy: Four genres", in *Philosophy in History*, ed. Richard Rorty, J. B. Schneewind, and Quentin Skinner (Cambridge: Cambridge University Press, 1984), 49–76, pp. 49–56.

This fissure between philosophy and history in the history of ideas was possible only in a definite time and place, namely, Britain in the 1950s and 1960s. Since the end of World War II, Britain lived in a more self-enclosed, isolated, and provincial world than ever before in its history. Forced self-reliance during the war became intellectual self-sufficiency after it. One of the foremost symptoms of that isolation was the virtual disappearance from the British intellectual horizon of the German historicist tradition, which had been a major force in Germany, and even in Britain and the US, throughout the nineteenth and early twentieth centuries. Had this tradition stayed more in public view, had it been more assimilated by English intellectuals, it is very unlikely that there would have ever been a divorce between philosophy and history in the history of ideas. For the historicist tradition not only questioned the kind of a-historical philosophizing presupposed by Strawson and Bennett, but it also stressed the importance of joining the historical and critical components of the history of philosophy. The great works of philosophical history of the historicist tradition—Rudolf Haym's *Herder* and *Die romantische Schule,* Kuno Fischer's *Geschichte der modernen Philosophie*, Eduard Zeller's *Geschichte der griechischen Philosophie*, Wilhelm Dilthey's *Leben Schleiermachers*, Ernst Cassirer's *Das Erkenntnisproblem in der Philosophie und Wissenschaft der neueren Zeit*—had managed to achieve something like a synthesis of the historical and critical in the history of philosophy. These works were exemplary, not only in explaining an author's intentions, context, and culture, but also in criticizing his main ideas; and because their criticisms came from within, *after* the most sympathetic reconstructions, they were all the more cogent and compelling.[14] Unfortunately, they remained untranslated and largely unknown. Lacking knowledge of such precedents, Strawson and Bennett were free to pursue philosophy with little or no history, and Skinner and Tully were able to do their history with little or no philosophy. The fissure into the historical and philosophical history of ideas was, therefore, an original and characteristic English phenomenon, one unheard of in Germany. For a German, the phenomenon was bound to appear "*unheimlich*".

The signs of the disappearance of the historicist tradition were everywhere in Britain in the late 1940s and early 1950s. Collingwood's *The Idea of History,* first published in 1946, said little about it, and what little it does say is hackneyed and inaccurate.[15] Popper's *The Poverty of Historicism* gave the British public the idea that historicism is the doctrine that history conforms to laws, and that it was the rationale for all the atrocities of totalitarian regimes.[16] Little did people realize that Popper's usage was eccentric, and that the term "*Historismus*" in Germany had for decades virtually the opposite meaning: namely, the doctrine that history has no laws but consists in unique events.[17] Last but not least there

[14] I am not claiming that all these works are the *non plus ultra* of historical scholarship and philosophical criticism, and that we should go back to read them today. They are exemplary chiefly in their general approach, in their attempt to join the philosophical and historical. Their importance today is mainly historical: they created the modern discipline we know as the history of ideas.

[15] R. G. Collingwood, *The Idea of History* (Oxford: Oxford University Press, 1946).

[16] Karl Popper, *The Poverty of Historicism* (London: Routledge, 1957).

[17] On the history of the term, see Dwight Lees and Robert Beck, "The Meaning of 'Historicism' ", *American Historical Review* 59 (1953–54), pp. 568–77; Gunter Scholtz, "Historizismus" in *Historisches Wörterbuch der Philosophie* (Stuttgart: Schwabe & Co., 1974), III, 1141–7; Christopher Thornhill, "Historicism" in the *Routledge Encyclopedia of Philosophy* (London: Routledge, 1998), IV, 443–6; Georg Iggers, "Historicism" in *Dictionary of the History of Ideas* (New York: Scribner, 1973), II, 456–64; Erich

was W. H. Walsh's very popular *The Philosophy of History: An Introduction*, which virtually jerrymandered the German historicist tradition out of existence.[18] Walsh claimed that the Germans were only interested in the speculative philosophy of history, rather than the logical analysis of historical method—though the exact opposite had been the case from Chladenius to Weber. And so it came to pass that the history of ideas had for a generation only one exemplar in Oxford: Isaiah Berlin.

The causes for the disappearance of historicism from the British intellectual horizon are all too understandable, having their roots in the tragic aftermath of two world wars. The historian of philosophy, however, has to learn to see beyond this, for the simple reason that the origins of his discipline lie within the historicist tradition. All philosophy benefits in knowing its roots, in going back to its past and seeing how it derives from a long history; and this is especially true for the history of philosophy itself. We can heal the rift between history and philosophy in the history of philosophy—so I will argue—only by rehabilitating classical historicist themes.

My chief aim here is to defend the history of ideas, as propagated by Lovejoy in the anglophone world and as first conceived by the German historicists. It has come under increasing criticism lately, from both the new historicists and philosophers. But it was, as originally conceived and largely practiced, what the history of philosophy should be: a synthesis of the historical and philosophical. Most criticisms of this tradition, as we shall see, have been based upon ignorance and misconception.

The division of the discipline into extreme historical and philosophical halves has been an ongoing disaster. It has produced anachronistic philosophy and antiquarian history. Richard Rorty recommends that the halves stay apart because they have such different methods and purposes;[19] but this is the problem, not the solution. Such a policy will only aggravate the shortcomings of each discipline when taken on its own. We will get archaic history that interests no philosopher, and anachronistic philosophical reconstructions that concern no historian. Philosophers will never get beyond their own contemporary intellectual horizons, and historians will treat the claim to truth in texts of the past as if they are no concern to us now. The ideal method of the history of philosophy should get beyond this division; it should be a synthesis of its historical and philosophical halves.[20]

Despite its disasterous consequences, the fissure of the disciple has had one redeeming value: we can see clearly the problems that result from dividing it, and the need to reconcile its two halves. We shall investigate each of these one-sided extremes in sections 3.1 and 4.1, showing how they generate problems when taken on their own. We shall then, in two short concluding sections, show how the old historicism can be compatible with the search for truth in the history of philosophy.

Rothacker, "Historismus", in *Schmollers Jahrbuch* 62 (1938), 388–99, and by the same author, "Das Wort 'Historismus' " *Zeitschrift für deutsche Wortforschung* 16 (1960), 3–6.

[18] W. H. Walsh, *The Philosophy of History: An Introduction* (London: Hutchinson & Co., 1951). A revised edition was published in 1958; by 1966 the book had gone through several printings. It was the introduction to the subject in Britain and the US for decades.

[19] Rorty, "Historiography of Philosophy", pp. 53, 55.

[20] The case for synthesizing the two forms was stated long ago by John Dunn, "The Identity of the History of Ideas", *Philosophy* 43 (1968), 85–104.

3. The Analytic Method and its Travails

The main aim of Strawson's and Bennett's approach to the history of philosophy, which we will call for short "the analytic method", has been to assess *the philosophical content* of a text. This method could be summarized in three basic rules. First, identify the problem behind a passage, the question the author is attempting to answer. Second, reconstruct the arguments the author gives for its solution, i.e., formulate them into premises and conclusion. Third, appraise these arguments, determining their formal validity and the quality of the evidence for them. This method reflects a certain view about philosophy itself: that it consists in problems and the attempt to solve them through syllogistic reasoning. The main reason to study the history of philosophy, we are told, is to see the strengths and weaknesses of different strategies for solving problems, so that we will then be in a better position to solve these problems ourselves. Thus the history of philosophy becomes something like a treasure chest of parables for contemporary philosophers.

As it stands, there is nothing necessarily a-historical or anti-historical about the analytic method. There are historians of philosophy who attempt to practice this method along with an historical one. They reconstruct the arguments behind a text with full knowledge of its background context and the author's intentions. However, as Strawson and Bennett practiced this method, it did become a-historical, indeed anti-historical. We need to recall that Strawson and Bennett themselves wanted to dispense as far as possible with historical scholarship. They believed that they could get to the essential meaning of the text, reconstruct its arguments with accuracy, and assess them with fairness, without having to engage in a study of an author's intentions and context. Thus Strawson, in the preface to *The Bounds of Sense*, admitted: "I have not been assiduous in studying the writings of Kant's lesser predecessors, his own minor works or the very numerous commentaries which two succeeding centuries have produced."[21] And Bennett warned in his *A Study of Spinoza's Ethics* that "delving into backgrounds is subject to a law of diminishing returns", and that we are "more likely to get his [Spinoza's] text straight by wrestling with it directly, given just a fair grasp of his immediate background."[22] Of Henry Wolfson's massive work to set Spinoza in his medieval context,[23] Bennett stated: "the labour and learning are awesome, but the *philosophical* proft is almost nil."

This a-historical practice was not gentlemanly idleness on Strawson's and Bennett's part, however much Oxford might be infected with that aristocratic affectation. Rather, it stemmed from a guiding hermeneutical assumption: that to understand a text is to formulate it in *our own* contemporary terms. If we must put a text in our terms, there is no need to understand it in the author's, so that historical scholarship becomes dispensable. Bennett stated this assumption perfectly explicitly: "To understand someone's thought you must get it into your own terms, terms that you understand. The only alternative is

[21] Strawson, *The Bounds of Sense*, p. 11.

[22] Bennett, *A Study of Spinoza's Ethics*, (Indianapolis: Hackett, 2984), p. 16.

[23] Henry Wolfson, *The Philosophy of Spinoza* (Cambridge, Mass: Harvard University Press, 1934), 2 vols.

to parrot his [the author's] words."[24] Strawson fully shared Bennett's approach, for in his review of *Kant's Analytic* he praised Bennett for making Kant our contemporary and for saving him from "the wrong kind of respect".[25] Strawson and Bennett had an unlikely ally in their approach: Richard Rorty, who claimed that to reconstruct philosophical significance of a text is to translate it into contemporary terms, into what it means for us with our current philosophical interests.[26]

While I accept the aim of the analytic method—the reconstruction and appraisal of arguments—as a necessary part of the philosophical enterprise, and while I also endorse the analytic method as formulated in the three rules above, I find Strawson's and Bennett's attempt to practice this method on its own without the aid of historical scholarship problematic. If our aim should be simply reconstructing and appraising arguments, as Strawson and Bennett insist, we are not likely to achieve these goals without historical knowledge of an author. The problem here is basic and insurmountable: that we cannot reproduce arguments with any accuracy, or evaluate them with any fairness, without a sound knowledge of their historical context and content. If we read a text on its own, we run the risk of missing key variables: the meaning of central terms, hidden presuppositions, underlying problems, how an author is responding to objections raised by contemporaries, and so on. Attempting to interpret a text by itself, apart from a knowledge of the other texts to which it responds, is like hearing only one fragment of a larger conversation, not knowing what other interlocuters have said and why they have said it. As Skinner put it, it is like attending a trial and only hearing the arguments for the defense.[27] The crucial point to see here is that knowledge of this context is crucial for understanding the *argument* of the text, its *philosophical* content; it is not simply a requirement for knowing its historical details and background. The philosophy is in the history, and we have to dig it out by careful historical excavation.

Attempting to skirt the need for historical scholarship by resorting to our own imagination, by consulting our intuitions about "what he [the author] could possibly mean", is a strategy of desperation. Such armchair a priori reconstructions work on the extravagent premise that our present cultural discourse and practices will be exhaustive of all logical possibilities. But there is no a priori reason to think that the boundaries of our contemporary imagination will be able to fathom the discourse and practices of another culture, which depend upon many factors, such as language, culture, and history, all of which we can know only a posteriori. It is indeed in just this respect that the historicist is likely to object that the analytic philosopher is only reading his own philosophical and cultural concerns into the products of past cultures. He will warn us that analytic reconstructions not based upon knowledge of historical context run the risk of ethnocentrism and anachronism.

The Strawson and Bennett approach is likely to work best on texts within our own culture and recent historical past. It breaks down completely, however, for texts written in very different epochs, in very different languages and very different cultures. We then face the

[24] Bennett, "Response to Garber and Rée", in *Doing Philosophy Historically*, ed. Peter Hare (Buffalo: Prometheus Books, 1988), pp. 66–7.

[25] Strawson, *Philosophical Review* 77 (1968), p. 332.

[26] Rorty, "The historiography of philosophy', pp. 52–3n1.

[27] The metaphor is in Skinner, 'A Reply to My Critics', *Meaning and Context*, p. 274.

perennial problems of translating, editing, and reconstructing often fragmentary or undecipherable texts. Strawson and Bennett presupposed that such work had already been done for them and that it was accurate and reliable. Having little knowledge of German, they relied entirely on Kemp Smith's translation of the *Kritik der reinen Vernunft*. Fortunate they were to have had such a translator. But what is the philosophical historian to do in the absence of translations? He has no recourse but to learn the language and to study the culture of the past. A division of labor between philosophical interpreters and historical translators and editors is not advisable for the all too simple and familiar reason that meaning depends on language and context. No discipline is more sensitive to the exact meaning of words than philosophy, and that exact meaning is something very individual, depending on a specific language and how it is used in a specific time and place. A philosopher needs to know this meaning, and he cannot rely on others to deliver it to him because translations always contain mistakes and they offer only one choice in the decipherment of meaning. This point is a commonplace for classical scholars, but it needs stressing for scholars of modern thought too. Paul Kristeller has declared unequivocally: "Anything written on Greek philosophy by a person who knows little or no Greek, whatever his other merits or gifts, has no validity or authority whatsoever."[28] The same holds, I would add, for writings on French and German philosophy.

All these criticisms, it might be replied, simply miss the point behind Strawson's and Bennett's practice. They want to make a historical work relevant to *our* contemporary philosophical culture, and they make no claim that this is what some historical figure actually meant. What matters is what he or she means to us, and not what he or she meant to his or her contemporaries in that past context. Bennett has expressed this point perfectly clearly by saying his attitude toward a text is entirely "*instrumental*", i.e., he wants to get something out of it for himself, making it serve his own ends and interests.[29] Rorty has come to his aid on this point by stressing that there is nothing wrong in principle with such instrumental practices "if they are conducted in full knowledge of their anachronism".[30]

There are two responses to be made to this defense. First, if we interpret texts simply by putting them in our contemporary terms, by making them useful for our purposes, we forfeit one of the great values of the history of philosophy: broadening our intellectual horizons by seeing different ways of conceptualizing the universe. One major point of doing the history of philosophy is getting beyond the limits of our conceptual age and learning that what seems natural and necessary might be only the result of our own cultural habits and prejudices. Philosophers too are victims of ethnocentrism, and their only means of escaping it is by acquiring the broader vistas of history. Second, philosophers who engage in recasting the past in contemporary terms rarely leave it at that: they almost always go on to claim that these terms are also the correct historical ones. Rorty is perfectly correct that philosophers who engage in these practices need to be fully self-conscious of their anachronism. The problem, however, is that such self-awareness is rare and not easily sustained. For the most part, these philosophers assume that they are explaining "what an author really

[28] Paul Kristeller, "Philosophy and its Historiography", *The Journal of Philosophy* 82 (1985), 618–25, p. 622.

[29] "Response to Garber and Rée", pp. 67–8.

[30] Rorty, "The historiography of philosophy", p. 53.

meant", that they are capturing "the spirit of his doctrines". There is rarely a complete disengagement from the historical object; one waits in vain for a confession that they are talking only about "what the author *ought to* have meant". Bennett, for example, does not claim to provide simply a creation re-interpretation of his historical figures but also a historical understanding of them; he claims to be trying "to get the author right, i.e., to find out what he actually meant."[31] We are told that putting an author in our own terms is the condition for *understanding* him, not only for making him relevant and useful for our concerns. Another example of someone caught in this confusion is John Rawls, who is at first very careful to distinguish the historical Kant from the Kant who anticipates his theory of justice; but even Rawls slips back into the language of talking about what Kant really meant.[32] These instrumental interpretations, or creative re-interpretations, if they are fully self-conscious and consistent, would have to detach themselves entirely from their alleged historical object. They would have to admit that they create a fiction useful for philosophical purposes but having no bearing on any historical person. To continue to attribute some vestige of historical reality to them would be to create a monster, a being that has both a normative and factual existence.

It is worthwhile to contrast the Bennett–Strawson *instrumental* conception of the history of philosophy with the *aesthetic* conception, which was once commonplace in the historicist tradition. According to this competing conception, each philosophical text should be treated like a work of art, i.e., as an end in itself and as having an organic unity where all parts fit together to form a whole. This conception is closely connected with the principle of interpretive charity, according to which we should seek coherence and unity in a work before criticizing it. Just as we should have sympathy for a work of art and must feel ourselves into its world, so, it is held, we should do the same for a work of philosophy. This metaphor gives us a very different reason why we should do the history of philosophy. Since these works are ends in themselves, just like art works, they have more than an instrumental value. We no more expect them to solve our current conceptual problems than we expect a Beethoven symphony to bring home our groceries.

Strawson, Bennett, and Skinner have complained about notions of interpretive sympathy, about attempts to seek coherence and unity in a text. They think that it blinds us to imperfections, that it finds coherence and order where there is often incoherence and disorder. Strawson and Bennett also object that this paradigm involves an attitude of reverence where criticism is more appropriate ("the right kind of respect"). All these warnings have a point. There are indeed limits to the organic metaphor, because works of philosophy are so often not like works of art: unlike Bach fugues, Goya portraits, and Wren buildings, they are terribly flawed, marred with inconsistencies, ambiguities, fallacies, and straightforward falsehoods. But we should not take this point too far. It is important to recall that the aesthetic conception was really intended as a *regulative* ideal: that we should proceed *as if* there were unity, as if the philosopher's work were a rational whole. Its critics have read it

[31] "Response to Garber and Rée", p. 67.

[32] See John Rawls, *A Theory of Justice*, Revised Edition (Cambridge, MA: Harvard University Press, 1999), pp. 221–7. Rawls is very careful to say that he is not providing an interpretation of Kant's texts; but he still warns against misunderstanding Kant and insists that Kant's "*main aim* is to deepen and to justify Rousseau's idea that liberty is acting according to laws that we give to ourselves." (p. 225; my italics)

as if it were a constitutive principle, a claim about the actual structure of a text. The purpose of the ideal was not to instill reverence but to make criticisms. It was assumed that such a principle of regulative charity, if pursued to its limits, is an aid to help us to discover irresolvable inconsistency and problems in a work; the reasoning behind it was that if apparent inconsistencies and ambiguities do not disappear after the most concerted attempt to find unity and clarity, then they really are irresolvable, true difficulties rather than apparent ones. This policy would thus assure us against premature and superficial criticisms; it assumed that the most compelling criticisms come *after* the deepest sympathy and charity. This point was another staple of the historicist tradition. Following it would have greatly improved Strawson's and Bennett's work, which has proven exasperating for most of its readers because it all too often shoots first and asks questions later; their criticisms are all too often premature, based on misunderstandings which a greater effort to take the text in its own terms and to interpret as a whole would have avoided.

4. Pitfalls of the New Historicism

The chief goal of the new historicism, as advanced by Skinner and Tully, has been to understand the *historical meaning* of a text. Its method too could be summarized in three basic rules. First, identify the historical context behind a text, which consists in two dimensions, one of which is intellectual, i.e., what has been written on the same topic by contemporaries and predecessors, and the other of which is social and political, i.e., the current crises, problems, and conflicts of the culture in which the author is writing. Second, recover the intention of the author, the specific purposes for which he or she wrote at a specific time and place. Third, attempt as far as possible to understand the text from the standpoint of the person who wrote it. The chief reason that we do the history of philosophy, these historians tell us, is to escape the intellectual confines of our own age, to broaden our intellectual horizons by grasping the values and beliefs of another age. Only in this way do we attain full self-knowledge and the ability to discriminate between the contingent and necessary in the sphere of culture.[33]

These rules are all perfectly sound and important on their own, and following them would go a long way toward removing some of the problems that arise in applying the analytic method in an unhistorical way. However, taken as a complete or sufficient guide in the history of ideas, they are inadequate and reduce the history of philosophy down to history alone.

The main problem with the new historicism is that it has virtually abandoned the question of truth. It is not simply that it has found this question irrelevant for the sake of historical understanding: it thinks that attempting to pursue truth through historical texts leads nowhere. It has two reasons for bracketing the whole question of truth. First, it fears that criticism of a historical text is bound to be question-begging, imposing the beliefs and values of our own time upon texts written in an age having different beliefs and values; in other words, criticism is bound to be anachronistic or ethnocentric. Second, it thinks that

[33] Skinner, "Meaning and Understanding", pp. 52–3.

once the meaning of a text has been sufficiently placed within its own context, then the question of truth becomes irrelevant, because the issues and arguments were important only at that time; they served ends peculiar to that historical context, and those ends have disappeared with history. Thus it is not of much interest to us now whether Hobbes' arguments for absolute sovereignty are valid, because, when placed in their proper seventeenth-century context, it becomes clear that all that was at stake was the powers of the monarch in seventeenth-century Britain.

That the new historicists have pushed the matter to this extreme there should not be the slightest doubt. They have been explicit and emphatic enough. Skinner, while repudiating the charge of relativism to his methods, tells us fairly and squarely that "the suggestion that we need to consider the truth of the beliefs under examination is I think likely to strike the historian as strange."[34] The main aim of the intellectual historian, he declares, is "to serve as a recording angel, not a hanging judge … to recover the past and place it before the present, without trying to employ the local and defeasible standards of the present as a way of praising or blaming the past."[35] And Tully has defended bracketing the question of truth on the grounds that we have no universal standards of right and wrong.[36]

However, there are still reasons for thinking that the search for truth is vital to the history of ideas. The problem of abandoning it is not only the risk of antiquarianism: to do so also violates the historicist's own rules by failing to honor an author's intentions. It is almost always a vital part of the author's intention, not only to respond to crises of his age, but also to address more general intellectual problems. Although his texts are very much the product of his context, they also make claims to truth about general issues which transcend the confines of his age. This is the case even for the most historically situated texts. Aristotle's arguments about slavery, for example, are surely only comprehensible in the light of the politics of the Greek city state around the fifth century BC; yet they also claim to be true about human beings as such and to be valid for all societies. Hobbes' theory of sovereignty, though a response to the civil conflicts of the 1640s, is still controversial because it pretends to outline an ideal government. We do not fully understand or appreciate Aristotle's or Hobbes' arguments if we limit them to their particular historical context, as if they were intended for their own age alone.

But insisting on making truth an intellectual objective of the history of philosophy raises a difficult question: How can the historian of philosophy criticize the past without lapsing into the very dangers of enthnocentrism and anachronism pointed out by the historicists? There was an old solution to this problem in the historicist tradition, one which is still valid today. This solution was worked out by Herder, Friedrich Schlegel, and Hegel in the late eighteenth and early nineteenth centuries, when they confronted a very similar problem in the wake of their own historicism. Refusing to drop the demand of criticism, yet still aware of the dangers of anachronism and ethnocentrism, Herder, Schlegel, and Hegel stressed the importance of an *internal* or *immanent* critique of a text. Attempting to criticize a text from an external standard, from some purported universal criterion of reason, they regarded with suspicion, because such standards and criteria all too often proved to be only

[34] Skinner, "A Reply to My Critics", p. 256.
[35] Skinner, *Machiavelli* (Oxford: Oxford University Press, 1981), p. 88.
[36] James Tully, "The pen is a mighty sword", in *Meaning and Context*, pp. 19–22.

the values of our age universalized for all mankind; applying them to a text and criticizing it through them was therefore bound to be question-begging. So all sound and secure criticism should be made according to the goals and standards of the author himself.

This solution still seems to me to be plausible and promising. It indeed provides us with our middle path between the current dilemma of "epistemology" and "hermeneutics".[37] Of course, the method of internal critique does not tell us which standpoint is true, nor does it provide us with positive evidence of any kind. All that it permits us to do is to reject a standpoint; in Popperian terms, it serves only as a means of falsification rather than verification. Still, it would be a mistake to underrate the value of such a tool. There are many aspects of a text, besides simple consistency, that can be taken into account with such an internal critique: whether the author has achieved *his own* goals; whether he gives sufficient evidence for his premises; whether his premises entail his conclusions; whether and how his meaning is ambiguous; and so on. All aspects of the formal structure of a text, as well as whether the author follows his intentions, are fair game for the critic.

It is important to do away with the common assumption, then, that all criticism somehow does violence to texts, as if it were always necessarily imposed by some alien hand. If we are skilled practitioners of internal criticism, then we have a right to say that texts simply criticize themselves, or, as Hegel put it, they "suffer violence at their own hands". Such internal criticism is not only possible but necessary, imposed by the very nature of a philosophical text. As Lovejoy, who always insisted on the value of criticism in the history of ideas, put it:

> When a man has given a reason for his belief, his moral approbation or disapprobation, his aesthetic preference, he is—happily or otherwise—caught in a trap; for the reason is likely to entail, or to seem to entail, consequences far beyond and, it may be, contrary to, the desire which generated it.[38]

Besides its forfeiture of criticism, there is another problem with the method of the new historicism. While its rules are perfectly sound as far as they go, we must not take them too far, as if they were sufficient for a full understanding of a text. Its rules seem to reduce understanding down to its *historical* dimensions, i.e., knowing the author's intentions and context. But the problem is that even if we knew everything about these historical factors, we would still have a weak understanding of a *philosophical* text. They leave out of account the *logical* dimension of a text, its internal formal structure. Such a structure involves what we might call its "logical geography", which comprises such factors as: 1) the premises necessary to justify a proposition; 2) the implications or consequences of a proposition; 3) the various meanings of central terms, 4) whether sufficient evidence is offered for a proposition; 5) internal contradictions, and so on. We say that we have understood a text only when we know this logical geography, only when we have learned to play with or manipulate its fundamental concepts to see what follows from them and what they depend upon. This is understanding in a deeper sense than simply knowing the historical context and intention,

[37] This dilemma is posed by Rorty in his *Philosophy and the Mirror of Nature* (Princeton, N.J.: Princeton University Press, 1979), pp. 315–94. That this is a false dilemma is fully recognized by Skinner, "A Reply to My Critics", *Meaning and Context*, p. 243.

[38] Lovejoy, "Reflections on the History of Ideas", p. 19.

for it involves the interpreters *active appropriation and assimilation* of the structure of the author's thought. Here we no longer learn but also explore.

In general, to understand and interpret a work involves knowing *both* dimensions, its innner structure and coherence, as well as its intention and context. If we admit that knowing the inner structure of a text, its logical geography, is central to its understanding, then it becomes difficult to separate the domains of interpretation and criticism, of historical explanation and philosophical truth. For in determining the logical structure of a text, we are also testing its concepts, stretching them to see what consequences they have and what basic facts they explain. In exploring that geography we inevitably come across problems, apparent inconsistencies and inadequacies, because we fail to find internal coherence and unity.

But if it is difficult to separate the tasks of historical understanding and philosophical criticism, then it is legitimate to demand that the historical method be complemented by the analytic one. For what the historicists have always demanded is that we have the deepest possible understanding of a text, that the task of understanding the past involves "trying as far as possible to think as they thought and to see things in their way".[39] But this entails not only knowing an author's intentions and context, but also knowing the logical geography of a work, grasping "the logical powers" of its central contentions, for these too determine how an author understands his world.

5. The Role of Intentions in the History of Philosophy

From all I have said so far in criticism of the analytic and historical methods, it follows that the best method of the history of philosophy is a synthesis of both. This synthesis would consist in reconstructing arguments on the basis of their historical context, and appraising them according to an internal critique. This would ensure that we are reconstructing arguments accurately, and that we are appraising them fairly, not begging questions against them by imposing standards of our own upon them. Rather than wasting our time and energy on bashing a straw man, we would at least know that we are engaging some actual historical figure.

Such a synthesis would keep alive a sense in which the history of philosophy is a search for truth, though, admittedly, it would secure this sense in a rather modest way. We would be able to determine what is true or false only within the context of a particular text, or only with respect to the views of a particular historical thinker. We would be far from having a knowledge of what is true or false on any general issue. Still, this is not a negligible gain, given that the internal critique of another philosopher is the first step in forming one's own philosophical views. I take it that this is what Bennett meant when he claimed to be learning philosophy from the criticism of a historical text. If this is so, his enterprise is sound and laudable.

[39] Skinner, "A Reply to My Critics", *Meaning and Context*, p. 252.

However, if we limit the history of philosophy to the synthesis of these methods as I have described them so far, something would still be missing from it. I have described this synthesis as if the chief goal of the history of philosophy were the reconstruction and appraisal of arguments. But this should not be the *sole* goal of the history of philosophy. Simply to reconstruct and appraise arguments is never sufficient in the history of philosophy for the simple reason that it alone does not tell us why someone put forward an argument in the first place. The historian of philosophy should not just be a logician who concerns himself with the validity of arguments given by the great dead philosophers in times past: he also needs to know why someone made these arguments—the purpose, motive or intention behind them. To this extent, Skinner is entirely correct in emphasizing the importance of an author's intention in historical understanding.

Why should we bother with knowing the author's intention, his motives in putting forward an argument? It is because by this means we can see why ideas and arguments are important to people, why they matter to them. This alone shows us what is at stake in intellectual discourse. To ignore intentions and motives, as if the only or essential concern of the historian of philosophy were with the truth of ideas and the validity of arguments, is naive. It is not a simple concern for the truth as such that motivates people to make arguments: as a matter of historical fact, they also have moral, political, and religious ends, which determine what they regard as the truth.

As soon as we add the concern with intentions and motives to the history of philosophy, it inevitably becomes a much more historical discipline. For to uncover these motives, we have to do detailed historical research. We need to know biographical facts about the author, including his or her religious beliefs or political allegiances, and we need to know how they were placed in the context of the religious, cultural, and political struggles of the day. A concern with these motives, and the historical facts necessary to understand them, had always been central to the historicist tradition, and it is not least in this respect that many of its works are still illuminating.

Some historians of philosophy, however, have stoutly resisted adding this dimension to their discipline.[40] To them, it seems to threaten the autonomy of philosophy, to sully its concern with the pure logic of arguments where questions of validity alone matter. The internal critique of a philosopher's work, they argue, should assess it on the basis of philosophical reasons alone, regardless of the author's motives in writing it. We should explain and assess a philosopher's views only in light of his conversation with other philosophers, these philosophical historians insist, and not go outside philosophical discourse, looking for non-philosophical motives, unless there is some kind of breakdown in the philosophical reasoning of the author which makes it impossible to understand him purely philosophically.[41]

But this conception of the autonomy of philosophy is a philosopher's fiction: it is false that philosophers are such a-historical beings, such purely rational beings concerned

[40] See Michael Frede, *Essays in Ancient Philosophy* (Minneapolis: University of Minnesota Press, 1987), xv–xvii; and his "The History of Philosophy as a Discipline", *The Journal of Philosophy* 85 (1988), 666–72, pp. 670–1. See also Jorge Gracia, *Philosophy and its History* (Albany: SUNY Press, 1992) pp. 221–34; and Calvin Normore, "Doxology and the History of Philosophy", *Canadian Journal of Philosophy*, 20 Supplementary Volume 16 (1990), 203–26, p. 222.

[41] Thus Frede, *Essays*, p. xv.

with only the truth as such; even when there are the best philosophical reasons for holding a position, a philosopher will usually have moral, religious, and political motives for propounding it. It should not be only in exceptional cases, when the philosophical argument breaks down, when we resort to explanation by intention: this should be a constant effort on the part of the historian of philosophy, even when a philosophical explanation is perfectly illuminating on its own. This conception of philosophy presents us with a false dilemma: either reasons or causes; and only when we can find no reasons should we resort to causes. But there are almost always reasons and causes. It is only in the most abstract and abstruse areas of metaphysics and logic that a philosopher's reasons are also likely to be the sole causes, though, often enough, even in these rarified realms we can usually trace a moral, religious, and political motive at work. No one should assume that metaphysics, just because it is abstract and arcane, escapes history. Descartes' *Meditations*, despite its a-historical pretensions, is a defense of the Roman Catholic Church; Berkeley's *Principles* are an apology for his High-Church politics against materialists, free-thinkers and skeptics; and Hegel's *Logik* grew out of his attempt to find reason in history, a position intended as a middle path between radicals and reactionaries.

Some philosophers complain that raising questions of intentions and motives confuse questions about the origins of a belief with questions about its justification. They commit nothing less than the "genetic fallacy", as if simply knowing the motives behind a proposition or argument were sufficient to discredit it. Appraising the motives is a very different business, they say, from assessing the argument. An argument put forward with the best motives can be fallacious; and an argument motivated by the basest motives can be formally valid.

These warnings about a genetic fallacy are well taken, but there is no reason why addressing the question of an author's intentions should fly in their face. We can investigate the motives behind a philosophy without claiming that these motives alone validate or discredit its ideas and arguments. The investigation of these motives should have for its purpose historical understanding rather than philosophical truth. It tells us why the philosopher made the argument, though it reveals nothing about its formal validity or truth.

That said, there is still an extent to which knowing the purpose or motive behind an argument does add an extra dimension to its appraisal. We can now ask whether the author achieves the purpose he or she intended through argument, whether it suffices for his or her ends. If the argument were true, would it fully establish what the author intended? Would it really support his or her cause? In assessing an argument in this way we are not committing the genetic fallacy, because it is one thing to ask whether an argument serves an author's ends and quite another to disregard it because one disapproves of his or her ends.

6. Historicism and Relativism

One might admit all these points about an author's intentions but still have qualms about a more historical approach to the history of philosophy, especially one that introduces questions about a philosopher's motives and historical context. The chief worry here is relativism. If a philosophy is the product of the author's moral, religious, or political motives, and

if morals, religion, and politics change over time, then there is the danger that philosophy too will become completely historical, that truth itself will depend on and vary with culture and epoch.

Judging from history, the danger of relativism is indeed not exaggerated. Some historicist thinkers, viz., Dilthey, came close to relativism, even if they did not expound it, and the relativism later proclaimed by Nietzsche and Spengler had its roots in the historicist tradition. The classic statement of historicist relativism is Hegel's famous dictum that every philosophy is the self-consciousness of its own age, that it is "its own time comprehended in thought."[42] It was because of statements like Hegel's that relativism became the chief objection to historicism for decades. Indeed, in the early decades of the twentieth century, historicism was virtually synomous with relativism.

Are these fears of relativism overstated? This raises a very large question that we cannot fully pursue here. But we can go some way toward resolving the anxiety if we examine the rationale for Hegel's statement and see whether it had the dire consequences so often attributed to it.

Part of the basis for Hegel's dictum was his social anthropology, one common to the historicist tradition. According to that social anthropology, the very identity of a human being depends on the society and culture of which he or she is a part; the needs,values and beliefs of each person are the product of education, which forms everyone according to the specific values and beliefs of a culture. This social anthropology was a reaction against the anthropology of the Enlightenment, according to which there is a universal human nature which remains the same in all cultures and all epochs, as if each individual has its fixed identity no matter the specific culture or epoch to which it belongs. If we hold that the identity of an individual is formed by society and culture, and if we also hold that societies and cultures change throughout time, then we must deny that individual identity is self-sufficient and permanent. We have to say instead that the individual is an essentially plastic, historical being. Its beliefs, values, and needs will be the product of its time and place.

Another part of the rationale for Hegel's dictum was a theory of meaning, which first appeared in Hamann and Herder, but which later became a common doctrine of the historicist tradition. Its classic expression appears in Wilhelm von Humboldt's essays on language. This theory tells us that the meaning of a word very much depends on the specific context and language in which it is embodied; we can grasp that meaning in its fullness only insofar as we know the language in which it is expressed and the culture in which it is situated. If we ask about the meaning of a word in any language, we have to ask ourselves how it is used by speakers and writers in that culture and language at that specific time and place. Through translation we can make equivalences to the languages and cultures of other people, but all translation is a form of abstraction, and often much is lost through it, viz., the very particularity and uniqueness that gives a word its full and concrete meaning. It is no accident that Hamann and Herder developed their theory in studying the poetry of different cultures, where the individual identity of words and phrases, their flavor and nuance, is most evident and important; while they hesitated to apply it to science, they were less wary about applying it

[42] Hegel, "Vorrede", *Grundlinien der Philosophie des Rechts* (Berlin: Nicolai, 1821), pp. xxi–xxii.

to philosophy, where the meaning of central concepts is also determined by culture and usage.

Both these premises—the social anthropology and theory of meaning—make philosophy very much a historical enterprise. Both what I say, and why I say it, will be historically conditioned and determined. The theory of meaning entails that what I say—the very content of my statements—will depend on my language and cultural context. What I mean will be determined both by the rules of my language at a particular time and place, and by my place in a conversation at a particular time and place, i.e, by what I intend to say in relation to past interlocuters. The social anthropology means *why* I say something also depends on the social and historical context. The central ideals behind a philosophy, the basic values that motivate it, are going to be historical because they will be the product of a specific culture and epoch.

Although philosophy is, in all these ways, deeply historical, it still does not follow that is entirely so and that all truth in philosophy must be relative. There is still an aspect of philosophy that is a-historical, which has to do with its logic, with the *relations* between propositions whatever the meaning of the propositions themselves. Given the meaning of the basic terms, however historically caused and constituted, when placed in propositions we can determine with universality and necessity their logical relation to other propositions. What a proposition implies, what it follows from, how it is compatible or incompatible with other propositions, are matters of sheer logic, and they will be valid for anyone who understands the meaning of its terms from the context in which they take place. Assuming that we know all the historically relevant facts that determine meaning, we can evaluate the strengths and weaknesses of the arguments of a philosopher, assured that our assessment is valid for everyone alike.

In his *A Study of Spinoza's Ethics* Jonathan Bennett tells us that he read Spinoza's great work because he wanted to learn philosophy from it, and because he hoped to discover "philosophical truth".[43] In the light of historicism this all sounds very naive. Yet, as we have seen, no historicist, no matter how relativist, can undermine the value and validity of an internal critique of a philosopher, provided that it is properly historically informed. It is at least in this limited and qualified sense that the search for truth has its place in the history of philosophy.

In the end, then, neither the introduction of an author's intentions nor the historical study of context should undermine the philosophical search for truth in the history of philosophy. The history of philosophy can be both history and philosophy in equal measure, and it can be pursued as a single discipline.

Of course, these are demands of a high order, because our historian of philosophy will be both philosopher and historian, a master at *Begriffsspiel* and a maestro of history. He or she will learn to endure with patience and compassion the criticisms of colleagues, who complain about veering too much toward history or too much toward philosophy. Above all, he or she will know how to revere and pay dues to the two gods of his discipline: content and context. Both these gods are very jealous; but for just that reason the historian of philosophy will always do best by serving them in equal measure.

[43] Bennett, *A Study of Spinoza's Ethics*, pp. 15, 35.

References

Beiser, Frederick, *The German Historicist Tradition.* Oxford: Oxford University Press, 2011.

Bennet, Jonathan, *Kant's Dialectic.* Cambridge: Cambridge University Press, 1974.

Bennet, Jonathan, *Kant's Analytic.* Cambridge: Cambridge University Press, 1966.

Bennet, Jonathan, *A Study of Spinoza's Ethics.* Indianapolis: Hackett, 1984.

Bennet, Jonathan, "Response to Garber and Rée", in *Doing Philosophy Historically,* ed. Peter Hare. Buffalo: Prometheus Books, 1988, 62–9.

Frede, Michael, "The History of Philosophy as a Discipline", *The Journal of Philosophy* 85 (1988), 666–72.

Frede, Michael, *Essays in Ancient Philosophy.* Minneapolis: University of Minnesota Press, 1987.

Collingwood, R.G., *The Idea of History.* Oxford: Oxford University Press, 1946.

Dunn, John, "The Identity of the History of Ideas", *Philosophy* 43 (1968), 85–104.

Gracia, Jorge, *Philosophy and its History.* Albany: SUNY Press, 1992.

Hegel, G. W. F., *Grundlinien der Philosophie des Rechts.* Berlin: Nicolai, 1821.

Iggers, Georg, "Historicism", in *Dictionary of the History of Ideas.* New York: Scribner, 1973. II, 456–64.

Kristeller, Paul, "Philosophy and its Historiography", *The Journal of Philosophy* 82 (1985), 618–25.

Lees, Dwight and Beck, Robert, "The Meaning of Historicism", *American Historical Review* 59 (1953–54), 568–77.

Lovejoy, Arthur, "Reflections on the History of Ideas", *The Journal of the History of Ideas* 1 (1940), 3–23.

Meier, H. G., "Begriffsgeschichte", *Historisches Wörterbuch der Philosophie,* ed. Joachim Ritter. Basel: Schwabe, 1971, I, 788–808.

Normore, Calvin, "Doxology and the History of Philosophy", *Canadian Journal of Philosophy* 20, Supplementary Volume 16 (1990), 203–26.

Popper, Karl, *The Poverty of Historicism.* London: Routledge, 1957.

Rawls, John, *A Theory of Justice.* Revised Edition. Cambridge, MA: Harvard University Press, 1999.

Rothacker, Erich, ed., *Archiv für Begriffsgeschichte.* Bonn: Bouvier, 1955.

Rothacker, Erich, "Historismus", *Schmollers Jahrbuch* 62 (1938), 388–99.

Rothacker, Erich, "Das Wort 'Historismus'", *Zeitschrift für deutsche Wortforschung* 16 (1960), 3–6.

Rorty, Richard, "The historiography of philosophy: Four genres", in *Philosophy in History*, ed. Richard Rorty, J. B. Schneewind and Quentin Skinner. Cambridge: Cambridge University Press, 1984. pp. 49–76.

Rorty, Richard, *Philosophy and the Mirror of Nature.* Princeton, N.J.: Princeton University Press, 1979.

Scholtz, Gunter, "Historizismus", in *Historisches Wörterbuch der Philosophie*. Stuttgart: Schwabe, 1974. III, 1141–7.

Skinner, Quentin, "Meaning and Understanding in the History of Ideas", *History and Theory* 8 (1969), 3–53.

Skinner, Quentin, "Social Meaning and the Explanation of Social Action", *Philosophy, Politics and Society* (fourth series), eds. Peter Laslett, W. G. Runcimann, and Quentin Skinner. Oxford: Blackwell, 1972. pp. 136–57.

Skinner, Quentin, "Motives, Intentions and the Interpretation of Texts", in *On Literary Intentions,* ed. D. Newton de Molina. Edinburgh: Edinburgh University Press, 1976. pp. 210–21.

Skinner, Quentin, "A Reply to My Critics", in *Meaning and Context,* ed. James Tully. Princeton: Princeton University Press, 1988. pp. 231–88.

Skinner, Quentin, *Machiavelli.* Oxford: Oxford University Press, 1981.

Strawson, Peter, The Bounds of Sense: An Essay on Kant's Critique of Pure Reason. London: Metheun, 1966.

Thornhill, Christopher, "Historicism", *Routledge Encyclopedia of Philosophy.* London: Routledge, 1988. IV, 443–6.

Tully, James, "The pen is a mighty sword", in *Meaning and Context,* ed. James Tully. Cambridge: Polity Press, 1988), 7–25.

Walsh, W. H., *The Philosophy of History: An Introduction.* London: Hutchinson & Co., 1951.

Wolfson, Henry, *The Philosophy of Spinoza.* Cambridge, MA: Harvard University Press, 1934. 2 vols.

CHAPTER 27

THE METHODOLOGY OF POLITICAL THEORY

CHRISTIAN LIST AND LAURA VALENTINI

1. Introduction

Political theory, sometimes also called "normative political theory", is a subfield of philosophy and political science that addresses conceptual, normative, and evaluative questions concerning politics and society, broadly construed.[1] Examples are: When is a society just? What does it mean for its members to be free? When is one distribution of goods socially preferable to another? What makes a political authority legitimate? How should we make collective decisions? What goals should our policies promote? How should we trade off different values, such as liberty, prosperity, and security, against one another? What do we owe, not just to our fellow citizens, but to people in the world at large? Is it permissible to buy natural resources from authoritarian governments? Can war ever be just?

Political theory is a long-established field. Its questions have animated thinkers since classical antiquity. Consequently, the methods of theorizing and substantive conclusions are diverse. In this article, we review the methodology of a core branch of contemporary political theory: the one commonly described as "analytic" political theory.

Given space constraints, we are not able to cover the history of political thought, the study of ideologies, the comparative study of political thought across cultures, and "continental" political theory, including "hermeneutic", "post-structuralist", and "post-modernist" approaches (for a more comprehensive discussion, see Leopold and Stears 2008). Nonetheless, the label "analytic" should not be interpreted narrowly. It is meant to refer to an argument-based and issue-oriented, rather than thinker-based and exegetical, approach that emphasizes logical rigour, terminological precision, and clear exposition. The term "analytic" is not intended to refer only to the logical and linguistic traditions of philosophy associated with

[1] We are most grateful to Kimberley Brownlee, Emanuela Ceva, Johannes Himmelreich, Mathias Koenig-Archibugi, Joseph Mazor, Florian Ostmann, Mike Otsuka, Miriam Ronzoni, and Kai Spiekermann for extensive and helpful written comments.

the Vienna Circle and philosophers in Oxford and Cambridge in the first half of the twentieth century. Contemporary analytic political theory goes well beyond conceptual analysis. At least since the publication of John Rawls's *Theory of Justice* (1971/1999), the development of normative *theories* has been one of the field's central concerns.

This review is structured as follows. In section 2.1, we briefly demarcate the scope of political theory. In section 3.1, we comment on the analysis of political concepts. In section 4.1, we introduce the notions of principles and theories, as distinct from concepts. In section 5.1, we discuss the methods of assessing such principles and theories, for the purpose of justifying or criticizing them. In section 6.1, we review a recent debate on how abstract and idealized political theory should be. In section 7.1, finally, we discuss the significance of disagreement in political theory.

One clarification about the nature of this review is needed. Although we cover established ground, we do so from an angle that will be somewhat unfamiliar to at least some political theorists—namely, an angle inspired by the philosophy of science. We have chosen this angle with a view to systematizing the activity of analytic political theorizing so as to make its connections with other fields of philosophy and positive science more transparent. This seems appropriate in the context of a handbook on philosophical methodology.

2. The Scope of Political Theory

To demarcate the scope of political theory, it is helpful to distinguish it from its most closely related neighbouring fields: political science, moral philosophy, legal theory, normative economics, and social ontology. We also offer some comments on the use of the label "political theory", as opposed to "political philosophy".

2.1 Political Theory and Political Science

Political theory can easily be distinguished from (positive) political science. Political science addresses empirical and positive questions concerning politics and society (for an overview, see Goodin 2009). It seeks to describe and explain actual political phenomena, such as elections and electoral systems, voter behaviour, political-opinion formation, legislative and governmental behaviour, the interaction between the legislative, executive, and judicial branches of the state, and the stability or instability of different forms of government. Political theory, by contrast, addresses conceptual, normative, and evaluative questions, such as what a democracy is, how we ought to organize our political systems, and how to evaluate the desirability of policies.

Political theory and political science can complement each other. Normative recommendations and evaluations of policies or institutional arrangements often rest on empirical premises. It is hard to arrive at a blueprint for a just society, for example, without understanding how society actually works, since normative recommendations may have to respect feasibility constraints (e.g. Gilabert and Lawford-Smith 2012). Thus political theory requires political science and the social sciences more generally. Similarly, when

political scientists investigate, for instance, whether democracy promotes economic development or whether free societies are more politically stable and less corrupt than unfree ones, they need to know what counts as a democracy or how to define freedom. These questions require the conceptual input of political theorists. Finally, among the large number of empirical questions that political scientists could investigate, some are undoubtedly more interesting, relevant, and pressing than others, and political theory can help shape the research agenda.

2.2 Political Theory and Moral Philosophy

While there is a natural division of labour between political theory and political science, the distinction between political theory and moral philosophy is subtler. Some scholars view political theory as a subfield of moral philosophy, in which the concepts and principles from moral philosophy are applied to political—and, more broadly, social—problems. If one views political theory in this way but also does not want it to collapse completely into moral philosophy, one must give a clear criterion of when a problem counts as "political".

Unfortunately, this is not straightforward. For instance, saying that a problem is political if it involves multiple people or their living together may seem plausible, but is too inclusive. Many problems in personal ethics, such as how to treat one's relatives or friends, would count as political on this criterion, even though they are not generally classified in this way. On the other hand, saying that a problem is political if it involves the state or government is too restrictive, because a number of problems outside the sphere of the state or government narrowly construed might still be seen as political.

Consider, for instance, gender relations in civic life or in the workplace, which many people regard as a political issue. The question of whether, and how, a distinction between the private or personal domain and the political or social one can be meaningfully drawn is controversial. Some feminist thinkers have famously challenged the possibility of drawing any such distinction and have endorsed the slogan "the personal is political" (for an overview, see Baehr 2013). In sum, if one wanted to distinguish political theory from moral philosophy by referring solely to the *substantive domain* of problems addressed, one could at best use some heuristic criterion to capture a conventional distinction, but this would yield no principled line.

Another way to distinguish political theory from moral philosophy is to invoke the *conditions of theorizing* in each field. The aim of moral theorizing, one might say, is to come up with the correct solution to any given moral problem *simpliciter*—the solution that, based on the theorist's comprehensive moral view, is right. The aim of political theory, one might argue, is different. The political theorist, at least under modern conditions, is engaged in problem solving under a particular constraint: the presence of pluralism and disagreement about how to solve the problem at hand (see Rawls 1996; Waldron 1999). Thus any compelling solution to the problems of political theory, such as how to define justice or how to design a legitimate procedure for collective decision-making, must appeal to people with a variety of (reasonable) viewpoints, precisely because those solutions are meant to apply to, and be acceptable in, pluralistic societies.

If we use such a methodological, rather than substantive, criterion for distinguishing political theory from moral philosophy, we need not worry about identifying a particular domain of problems that counts as political. Rather, we can say that the hallmark of political theory is its mode of theorizing, against the background of (reasonable) pluralism. We return to this idea in section 7.1, where we discuss the significance of disagreement in political theory. Of course, substantive and methodological criteria for distinguishing political theory from moral philosophy can be combined.

2.3 Political Theory and Legal Theory

We now turn to the relationship between political theory and legal theory. The two overlap, and it is difficult to draw a sharp distinction between them. We might again arrive at a rough distinction by using some heuristic criterion to identify what counts as "legal" or "related to the law". As with the attempt to distinguish the "political" from the "private" or "personal", however, we cannot expect any such criterion to yield a definitive distinction.

Alternatively, we might try to distinguish political theory from legal theory by identifying different modes of theorizing associated with each field. For instance, an argument to the effect that justice *simpliciter* requires respect for human rights and certain universal welfare protections is distinct from an argument to the effect that a particular constitution or kind of legal system, properly interpreted, requires them. One might say that the former argument belongs to political theory, the latter to legal theory (see the discussion of the nature of legal interpretation in Dworkin 1986).

More generally, one might say that the deontic concepts used in legal theory, such as *legal* permissibility, *legal* rights, and *legal* obligations, are different from their counterparts in political or moral theory and therefore require a different analysis.[2] Still, political theory and legal theory are best seen as overlapping fields of enquiry.

2.4 Political Theory and Normative Economics

Political theory also overlaps with normative economics, especially with social choice and welfare theory. Social choice and welfare theory is the formal, but also normative and evaluative, study of (i) collective decision-making procedures, (ii) mechanisms for allocating benefits and burdens in society, and (iii) methods by which a social planner, policy maker, or institutional designer can assess the goodness or desirability of different social states, policies, or institutions. Normative economists investigate these—(i), (ii), and (iii)—by introducing desiderata that any reasonable procedure, mechanism, or method is required to satisfy, and then asking which procedures, mechanisms, or methods, if any, satisfy the given desiderata. (For a survey, see List 2013.)

[2] Of course, even within legal theory, one can distinguish between what is legally permissible according to a specific set of laws (e.g. those of a particular country) and what the standard of legal permissibility should be.

The substantive questions addressed in social choice and welfare theory are similar to some of those addressed in political theory. For example, questions such as how to distribute benefits and burdens are addressed by political theorists and normative economists alike. Indeed, Rawls's *Theory of Justice* was, in part, influenced by the normative works of economists such as Kenneth Arrow (1951), John Harsanyi (1955), and Amartya Sen (1970). Similarly, normative economists frequently draw on moral and political theory. For example, John Roemer's (1998) formal work on distributive justice is influenced by G. A. Cohen's (1995) work in political theory; and the work on variable-population social choice by Charles Blackorby, Walter Bossert, and David Donaldson (2005) is influenced by Derek Parfit's (1984) work on population ethics.

Arguably, the main difference between political theory and social choice and welfare theory is not a substantive one (although the former is broader than the latter), but a methodological one. Mainstream political theory is a non-formal discipline, making at most limited use of formal methods from mathematics, logic, and economics, while social choice and welfare theory is predominantly formal.

2.5 Political Theory and Social Ontology

A less well-known but growing field in the neighbourhood of political theory is social ontology. Social ontology investigates the nature of phenomena such as joint intentions, collective actions, social norms and conventions, group agency, and institutions (e.g. Gilbert 1989; Pettit 1993; Searle 1995; Bratman 1999; Tollefsen 2002; Tuomela 2007; List and Pettit 2011). Its central question, roughly speaking, is: What are the building blocks of the social world, and how are they related to one another, to the individuals involved, and to the physical world?

The substantive questions in social ontology are distinct from those in political theory and in some respects prior to them. Social ontology is primarily a positive and explanatory field rather than a normative or evaluative one. Its relevance to political theory lies in the fact that political theory cannot get off the ground unless we are clear about what entities and properties exist in the social world. For example, before we can answer questions about rights, obligations, and responsibilities, we need to know what entities can be the bearers of rights, obligations, and responsibilities. Are these restricted to individuals, or do they also include certain collectives? Should we regard the state as an agent, as a mere collection of individuals, or as some kind of special fiction? Similarly, before we can answer the question of what is or is not socially desirable, we need to know what the possible objects of value might be. It is difficult to determine, for instance, whether there should be any cultural rights or state subsidies for cultural activities unless we can give at least a partial answer to the question of what we mean by "culture".

Since different social-ontological theories give us different accounts of what entities and properties there are in the social world, they can, in turn, impose constraints on what the possible objects of analysis in political theory might be.

2.6 Political Theory and Political Philosophy

Although some scholars distinguish between political *theory* and political *philosophy*, this is mainly a conventional distinction. It refers, roughly, to the different styles of political theory conducted in political science and philosophy departments, respectively, especially in North America. (In the UK, much of what is conventionally called "political philosophy" is traditionally conducted in political science departments.) Arguably, "political theory" is the slightly more inclusive and interdisciplinary label, referring not only to philosophical work but also to a variety of other approaches. As mentioned, here we focus specifically on the analytic branch of political theory.

3. Concepts in Political Theory

A long-standing concern in political theory is the analysis of political concepts: *freedom, equality, justice, authority, legitimacy, democracy, welfare*, and so on. Each of these has been interpreted and defined in numerous ways, and political theory can help us clarify the advantages and disadvantages of different interpretations and definitions. The bulk of political theory in the decades prior to the publication of Rawls's *A Theory of Justice* was conceptual analysis.[3]

The analysis of political concepts is relevant, not only to normative theory building (e.g. any *theory* of liberalism must employ some *concept* of liberty) but also, as already noted, to positive work in political science. Thus conceptual analysis is, in some sense, the least normative or evaluative part of political theory.

Of course, when we analyse concepts such as *freedom* and *democracy*, we are usually interested in the kinds of freedom and democracy that we find valuable or normatively required. Hence the ultimate motivation for our analysis may well be a normative or evaluative one: we may wish to clarify these concepts in order to express normative or evaluative principles in terms of them. Logically, however, the question of how to understand *freedom* and *democracy* is distinct from the question of whether freedom and democracy so understood are valuable. Indeed, political scientists may sometimes be interested primarily in whether *freedom* or *democracy* can serve as independent variables in explanations of political phenomena (e.g. when they investigate whether freedom promotes prosperity or whether democracies are less prone to fighting wars against one another), irrespective of any considerations of value.[4] In this section, we introduce some key ideas relevant to the analysis of political concepts.

[3] For a classic work on political concepts, see Oppenheim (1981).

[4] Of course, political scientists may also consider *freedom* or *democracy* as dependent variables, such as when they ask which social and economic conditions promote each. Think, for instance, of the literature on the transition to, and consolidation of, democracy (e.g. Linz and Stepan 1996). In such studies, considerations of value may plausibly affect our choice of interpretation of the relevant concepts.

3.1 What is a Concept?

We use concepts to categorize or classify objects.[5] The concept *democracy*, for example, may help us distinguish between those forms of government that are democratic and those that are not. The concept *legitimacy*, in relation to acts of state coercion, may help us distinguish between those acts of state coercion that are legitimate and those that are not.

For the purposes of this article, we assume that any concept has a *domain of application*. This is the set of objects of which it is meaningful to ask whether they fall under the given concept or not. We might say, for example, that the domain of the concept *democracy* is the set of all systems of government or decision-making. For any object in that set—that is, any system of government or decision-making—we can meaningfully ask whether it is democratic or not. By contrast, for objects outside that domain, it is not meaningful to ask whether they are democratic or not. It makes no sense to ask, for instance, whether an equilateral triangle or a mountain is democratic or not (cf. Dworkin 1986, 75). Note that the domain of the concept *democracy*—or *justice*, or *freedom*, and so on—can be variously specified; we will return to that point.

Further, any "classically well-behaved" concept has *defining conditions*. These determine, for any object in the concept's domain, whether that object falls under the concept ("satisfies it") or not. In our example, the question of whether a particular system of government or decision-making, say the political system of Iceland, is democratic or not depends on a variety of features of that system: how decisions are made, who participates in those decisions, how the participants provide their input, how the decisions are implemented and by whom, and so on.

Philosophers are divided over the extent to which concepts in general have defining conditions (for a review, see e.g. Margolis and Laurence 2012). Many of our commonsense concepts arguably lack such conditions. We may be able to pick out some paradigm instances (or "prototypes") of *redness* or *beauty*, but may be unable to arrive at clear-cut necessary and sufficient conditions that an object must satisfy to be red or beautiful (for a related discussion, see Dworkin 1986, ch. 2). By contrast, in theoretical work, it is usually desirable to look for concepts with defining conditions. Occasionally, however, some theoretical concepts may be regarded as undefined "primitives" or as characterizable only through "prototypes".

Finally, a concept's *extension* is the subset of the domain consisting of precisely those objects that fall under the concept ("that satisfy it"). If the concept has defining conditions, these determine the extension. The extension of the concept *democracy* is the set of all those systems of government or decision-making that, according to the concept, count as democratic. More generally, there can be "non-binary" concepts, which do not subdivide objects into those falling under the concept and those not falling under it, but which

[5] We here cannot discuss the ontological status of concepts, on which there are several rival views in philosophy. Instead, we rely on a relatively simple characterization of concepts, emphasizing the fact that we use concepts to categorize or classify objects and that they serve as ingredients in the activity of political theorizing. For a broader discussion of concepts as the constituents of thoughts, see Margolis and Laurence (2012). For a cognitive-science treatment of concepts as locations or regions within "conceptual spaces", see Gärdenfors (2000).

instead classify objects on one or several dimensions that may each admit of degrees. For example, *equality* and *welfare* are non-binary concepts. The level of equality or inequality in a particular distribution of goods is a matter of degree, and different interpretations of *equality* give us different accounts of when one distribution counts as more equal than another (see e.g. Sen 1980). Similarly, a person's welfare is a matter of degree and, on some accounts, even given by a vector of multiple attributes, representing different dimensions or aspects of welfare. Sen (1987) has argued for the "constitutive plurality" of the concept *standard of living*: there are multiple dimensions on which a person's standard of living can be categorized.

For practically any salient concept in political theory, there are debates about what the domain of application is, what the defining conditions, if any, are, and which objects belong to the concept's extension and which do not. Just think of the many different ways in which the concept *democracy* may be understood. The domain of application may be specified in a variety of ways: for example, as a set of voting procedures, as a set of decisions, or as a set of entire political systems. Similarly, we may be divided over the defining conditions that determine whether something is democratic or not: for example, do only formal, constitutional features of the political system matter, or are features of actual political practices relevant as well, and if so, how? And is democracy precisely definable at all? In consequence, we may end up with different extensions of the concept *democracy* (see e.g. Christiano 2008; List 2011). Similar considerations also hold for other key political concepts, such as *justice, freedom, equality*, and *legitimacy*. Indeed, many political concepts are what political theorists call "essentially contested" (Gallie 1955).

3.2 Concepts and Conceptions

Political theorists, following Rawls, who in turn follows H. L. A. Hart, sometimes distinguish between *concepts* and *conceptions* (Rawls 1971/1999, 5). Concepts, in that terminology, are less fully specified than conceptions. For example, we may have a broadly outlined *concept* of freedom as the absence of constraints on agents' actions, which still leaves open what kinds of constraints, agents, and actions matter (MacCallum 1967). A full specification of those constraints, agents, and actions yields a precise *conception* of freedom. Different conceptions can thus be compatible with the same broad concept.

We can translate the distinction between concepts and conceptions into our earlier terminology by defining a *conception* exactly as we defined a concept in the last section, where the domain, defining conditions (if any), and extension are fully specified, and redefining a *concept* as a broader family of such conceptions, with some aspects of the domain, defining conditions, or extension left open.

3.3 Desiderata on Concepts

The following is a list of desiderata that systematize requirements often implicitly employed by political theorists engaging in conceptual analysis:

Respecting our intuitions: We may want to interpret a concept, such as *freedom* or *democracy*, in a way that is broadly in line with our intuitions, especially if this concept has a common-sense interpretation. If we arrived at an interpretation of *freedom* that classifies some intuitively clear cases of unfreedom as instances of freedom, or vice versa, this would be suspect. For example, G. A. Cohen has famously criticized Robert Nozick's "moralized" conception of freedom (according to which freedom is, roughly, the absence of rights-violating interference) on the grounds that it delivers counterintuitive judgements, such as that a *justly* imprisoned criminal is not unfree (G. A. Cohen 1988). Similarly, we would be disturbed if our interpretation of *democracy* classified some clear-cut cases of tyranny as democratic. In such cases, we would either have to give strong reasons for overruling our intuitive judgements or search for a better interpretation of the concept. Later, we discuss Rawls's method of *reflective equilibrium*, which may help us adjudicate cases in which our theoretical conclusions conflict with our intuitive judgments (Rawls 1971/1999; Daniels 2013).

Playing the right normative, evaluative, or descriptive role: We may be interested in a concept because we wish to use it in some normative or evaluative principle or in some explanation in political science. A good interpretation of the concept is one that successfully plays the intended normative, evaluative, or descriptive role. Different roles might require different interpretations of the concept. For example, if we want the concept of justice to offer a comprehensive picture of how society should be organized, we are likely to develop a thicker account of it than if we understand justice as one value among the many that should guide institutional design (Rawls 1971/1999; G. A. Cohen 2008, 271–2). However, to avoid a proliferation of rival interpretations of the same concept, we might also be looking for a single interpretation that can successfully play multiple roles.

Standing in the right relationship to other concepts: Since we typically employ concepts not in isolation but in connection with other concepts, we may require these concepts to be related to each other in the right way. If we take rights to entail obligations, for example, our interpretation of *rights* may constrain the way in which we can consistently interpret *obligations*. Sometimes we may wish some concepts to be directly inter-definable. It is often held, for instance, that the concept *permissibility* must be definable in terms of the corresponding concept *obligatoriness* and vice versa: it is permissible that *p* if and only if it is not obligatory that *not p*. Consequently, our joint analysis of *permissibility* and *obligatoriness* must respect this constraint. Finally, we may wish some given concepts to be sufficiently "differentiated" from one another, in order to avoid redundancies or confusions. (For a related discussion of concept formation in the social sciences, see Gerring 1999.)

Having defining conditions that are neither too "thick" nor too "thin": Even when we have settled the domain of a given concept (e.g. we focus on a concept of freedom whose domain is a set of acts, as opposed to a set of agents or institutional arrangements), we might still be divided over the concept's defining conditions. In the case of *freedom*, a huge variety of different defining conditions have been proposed (for overviews, see Carter 2012; Lovett 2013). To narrow down the range of possibilities, we might require that the defining conditions be neither too "thick", nor too "thin": they should not refer to any "irrelevant" facts about the objects to be categorized, but refer to all "relevant" facts. There can then be debates about which facts are or are not relevant and what counts as too "thick" or too "thin". For example, a concept is *moralized* if its defining conditions refer to some normative or evaluative

facts. A concept is *non-moralized* otherwise.[6] Nozick's concept of freedom as the absence of rights-violating interference, which we have mentioned earlier, is moralized in this sense. Political theorists are often divided over which concepts in political theory should be moralized. Similarly, a concept is *modally demanding* if its defining conditions refer not only to facts about the actual world but also to facts about other possible worlds—that is, facts about what would be or might be, not merely about what is. A concept is *modally undemanding* otherwise. Pettit, for instance, argues that *freedom* is modally demanding: a slave with a benevolent, non-interfering master still counts as unfree, because there is a nearby possible world in which the master interferes (Pettit 1997). Similarly, concepts such as *security* and *peace* are arguably modally demanding, referring not only to the absence of relevant harmful actions or military conflicts in the actual world but also their continued absence in a range of nearby possible worlds. Political theorists are divided over which, if any, concepts should be modally demanding (see e.g. Pettit 2011; Southwood 2013).

Having defining conditions that are epistemically accessible: Depending on the intended use of a concept, we may require its defining conditions to be such that it is possible, at least in principle, for us to *know* whether an object meets them. For example, a concept of welfare whose defining conditions refer to certain kinds of mental states that are in principle inaccessible to any observer would be of little practical use. Similarly, the defining conditions of *justice* under Robert Nozick's account, which refer to the entire history of transactions leading to the current distribution of entitlements, fail to meet this desideratum and thus render Nozick's account somewhat practically inert (Nozick 1974). Of course, the context and intended use may determine what counts as epistemically accessible.

3.4 What Concepts are Not

Concepts should not be confused with principles or theories—the topic of sections 4.1 and 5.1. In particular, principles and theories have propositional content and may be true or false, while concepts, by themselves, cannot be true or false. They only categorize objects and, in doing so, can be more or less useful, more or less plausible, and more or less in line with established use or with our considered judgements.[7] To give a simple illustration from outside political theory, the concepts *red, green*, or *triangular* are neither true nor false. Only statements in which they occur can have truth-values, such as "tomatoes are red" (true), "snow is green" (false), and "rectangles are triangular" (false).

Even if we have a full account of when a political system is democratic or what it means for someone to be free, this still leaves open the question of whether democracy or freedom are desirable and whether we ought to promote each of them. We need principles or theories—making statements such as "we ought to respect freedom" or "we ought to make decisions democratically"—to address the latter questions. Such principles or theories are then capable of being true or false. Simply put, concepts can serve as building blocks of principles, which can serve as building blocks of theories.

[6] The definition of moralized and non-moralized concepts requires suitable adjustments if there are no evaluative or normative facts.

[7] Recall that the defining conditions of any concept simply specify when an object falls under the concept, not whether the concept is true or false.

Still, people sometimes say things such as "freedom as non-interference is the true conception of freedom". Assertions of this kind are best interpreted as abbreviations for claims such as "freedom as non-interference is the conception of freedom that is, in some relevant sense, most appropriate", which, in turn, could mean that it best captures our established use of *freedom* or alternatively our considered judgments about what counts as free. Literally, however, the claim that one concept of freedom is the true one is not meaningful, since only things with propositional content can have truth-values. Concepts do not have propositional content: the extension of a concept is not a set of possible worlds (which is what the extension of a proposition is), but a set of objects.[8]

Even on its charitable interpretation in terms of appropriateness, the assertion "freedom as non-interference is the true conception of freedom" is ambiguous. As noted, it would have to be understood as implying that, among the many different interpretations of *freedom*, one stands out as "most appropriate". But, unless we specify a criterion of appropriateness, there is no unique such interpretation. For example, one interpretation may be most in line with our common-sense use of the word "freedom", another most in line with our considered judgements about what counts as free, a third most suitable for playing a particular role in a theory of justice. Different criteria of appropriateness may diverge, and there is no application-independent criterion. Thus expressions such as "the true conception of freedom" are, at best, shorthand for more elaborate expressions involving criteria of appropriateness. To avoid ambiguity, it is best to spell those out explicitly.

4. Principles and Theories in Political Theory

While analytic political theory until the early 1970s was primarily concerned with the analysis of concepts, John Rawls's *Theory of Justice* (1971/1999) invigorated the quest for theories and the principles underlying them. Rawls formulated some principles of justice, which are the basis of his theory of how we should organize the "basic structure of society", namely its main political, legal, and economic institutions.

In this section, we discuss the notions of principles and theories. Although these are widely used in political theory, they are seldom carefully defined. We hope, therefore, that our discussion will be clarifying.

4.1 What is a Principle?

A *principle* is a statement—a proposition expressed in language—that applies, at least potentially, to more than one case. Usually, this is marked by the occurrence of expressions with appropriate quantifiers, such as "for all X, subject to certain conditions, Y is the case". A principle is *evaluative* if it has evaluative content: for instance, it includes evaluative

[8] Note that the extension of a statement or proposition is the set of those possible worlds in which the statement or proposition is true.

predicates or concepts such as *good* or *bad, better* or *worse, desirable* or *undesirable*. A principle is *normative* if it has normative content: for instance, it includes deontic operators such as *ought, may, permissible, obligatory, right*, or *wrong*. A principle without any evaluative or normative content is *positive*. Such principles are common in the sciences; think of the principle of conservation of energy in physics.

Classic examples of normative principles are the Ten Commandments from the Bible, the Golden Rule ("You should treat other people in the way in which you would like them to treat you"), Bentham's principle of utility ("An action is right if it maximizes total utility"), and more recently, Rawls's principles of justice. Roughly speaking, these state that each person is entitled to the most extensive system of individual liberties, compatible with a similar system for everyone else (the "equal-liberty principle"); and socio-economic inequalities are permissible if and only if they are compatible with a system of fair equal opportunities (the "fair equality of opportunity principle") and benefit the least well-off members of society (the "difference principle") (Rawls 1971/1999).

The *propositional content* of a principle is the set of all its implications. Sometimes this propositional content may depend on auxiliary assumptions. As already noted, their having propositional content sets principles apart from concepts, which merely offer categorizations.

4.2 What is a Theory?

The word "theory" is commonly used in two distinct senses. First, it can refer to an entire academic field or area of enquiry, such as when we speak of "political theory" or "economic theory" as general areas to which curricula or scholarly journals are devoted. Second, the word can refer to a specific theory within such an area, such as Rawls's theory of justice, the theory of the firm in economics, or Newton's theory of physics. Our focus here is on theories in this second, specific sense. Surprisingly, there exists no canonical definition of a theory in that sense in political theory.

To provide a starting point, we propose a simple definition inspired by the philosophy of science but adapted to the present context.[9] We define a *theory* as a set of statements—propositions expressed in language—which is a candidate for playing some theoretical or practical role and which is, ideally, representable as the set of all implications of some underlying principles. The set of principles from which the theory can be derived—if there is such a set—is called the *theory formulation*.

Although loose and abstract, this definition has some merits. First, it allows us to view positive theories (in the sciences) and normative or evaluative theories (in moral philosophy or political theory) as instances of the same general category. Second, it makes transparent the differences between them. For example, the roles played by theories can range

[9] We here follow broadly what is often called the *syntactic* approach to defining theories (where a theory is defined as a set of sentences/propositions with certain properties); it is arguably the most conventional approach. For a classic exposition, see Quine (1975). It is also worth exploring the rival *semantic* approach (where a theory is defined as a set of models with certain properties), but given space constraints, we set this aside here (see van Fraassen 1980). For an introduction to the philosophy of science, see Okasha (2002).

from descriptive, explanatory, and predictive (in the case of positive theories) to evaluative and prescriptive (in the case of evaluative or normative theories). Third, the definition allows us to identify the special challenges that arise when we construct and assess normative or evaluative theories.

Paradigmatic examples of theories according to our definition are Newton's theory of physics and Rawls's theory of justice. Each can be viewed as a set of statements, entailed by some underlying principles, which can play a descriptive, explanatory, or prescriptive role. Newton's theory is the set of all statements entailed by Newton's principles of physics, perhaps together with some empirical premises about the solar system or other physical systems of interest. It can be used, for instance, to explain and predict the trajectory of the planets around the sun and to guide such engineering projects as travelling to the moon and safely back. Rawls's theory is the set of all statements entailed by Rawls's principles of justice, perhaps together with some empirical premises about relevant social conditions. It can play a prescriptive or normative role, guiding us in the design of social institutions (for an earlier discussion of the relationship between normative and positive theories, see McDermott 2008).

For our purposes, the biggest structural difference between Newton's and Rawls's theories is that one is positive and the other normative. Indeed, the principles underlying Newton's theory are positive principles, while those underlying Rawls's are normative ones. Generally, a theory is *positive* if it has no evaluative or normative content; it is *evaluative* or *normative* if it has such content. Evaluative theories that offer evaluations of "goodness" or "betterness" are also called *axiological*.

Earlier, we associated evaluative content with the occurrence of evaluative predicates or concepts, such as *good, better, desirable*, and so on, and normative content with the occurrence of deontic operators, such as *ought, may*, and so on. Different accounts of what qualifies as evaluative or normative content can be given, and we need not commit ourselves to one such account here. On any reasonable account, Rawls's theory will come out as normative and Newton's theory as positive.

While Newton's and Rawls's theories are paradigmatic instances of our definition, a theory need not be self-consciously theoretical. A set of rough and informal principles describing how ordinary objects behave when pushed, dropped, or thrown can constitute a "folk" theory of motion that is predictively useful in everyday contexts. Similarly, a set of basic principles describing how animals respond to noise, movement, and the presence of humans may constitute a simple predictive theory of animal behaviour that members of hunter–gatherer societies might have used to guide their actions. We also routinely employ normative theories without self-consciously doing so. For example, a set of simple principles specifying how we should or should not treat others may constitute a simple "folk" theory of personal ethics.

4.3 The Distinction Between a Theory and the Support for It

Our simple definition of a theory is silent on whether the theory is true, useful, or good in some sense. A theory that is false, irrelevant, or superseded still counts as a theory. Describing something as a theory carries no assessment of its truth or acceptability. Thus

the most far-fetched and implausible conspiracy theory—for example, in science or in history—still qualifies as a theory.

Similarly, the attempt to downgrade a set of statements by asserting "it is *only* a theory", as critics of evolutionary theory or global-warming sceptics sometimes do when they describe their target, makes little sense. Calling something a theory is consistent with its being well supported and true, just as it is consistent with its being speculative and even false.

An important distinction is that between a theory and what is offered in its support. A theory, as we have defined it, is distinct from any arguments, evidence, or justification given for it. For example, Newton's actual or hypothetical experiments—such as how an apple fell onto his head (though supposedly a myth)—are not part of his physical theory itself: they are part of the evidence he had for that theory. Similarly, Rawls's arguments in support of his principles of justice, such as his original-position thought experiment (a hypothetical choice situation in which the parties to the social contract must agree on principles governing the basic structure of society), are not strictly speaking part of his theory of justice itself. The original-position thought experiment, like any physical experiment or any social scientist's observation, is offered in support of the theory in question. Thus Rawls's book, *A Theory of Justice*, consists of the theory itself (everything entailed by his principles of justice), together with arguments in support of it (most notably, the original-position thought experiment and reflective-equilibrium considerations), comparisons with rival theories (such as utilitarianism), and a fair amount of commentary on the theory's interpretation and its applications. This distinction, between a theory and the argument or evidence in support of it, will help us clarify some recent methodological debates in political theory, about abstraction and idealization, which we discuss in Section 6.

An important task, separate from defining a theory, is therefore to identify the requirements for a good, or acceptable, theory, and to spell out how we can assess it.

5. The Assessment of Principles and Theories

There are two kinds of criteria that we may use to assess—especially to justify or to criticize—principles and theories: "internal" and "external" criteria. The former concern the way the principles or theory are formulated and their internal logical structure. Criteria such as consistency and parsimony fall into this category. The latter concern the relationship between the principles or theory and what these are "about": their normative or evaluative content, in analogy with the empirical or descriptive content of a scientific theory. Criteria such as truth or normative adequacy (in analogy with truth or empirical adequacy in science) fall into this second category. In this section, we discuss the two kinds of criteria ("internal" and "external") in turn.

Throughout this discussion, we focus on theories, rather than principles, as the units of assessment. This is no loss of generality. We are usually interested, not in individual principles in isolation, but in sets of principles that we wish to assess together. If we wish to assess a principle by itself, we can view it as a special case of a theory, namely a theory that consists just of the principle and its implications.

5.1 Internal Criteria

Although we have kept our definition of a theory deliberately thin, defining it simply as a set of statements that may play some theoretical or practical role and that is, ideally, derivable from some underlying principles, we usually want theories to satisfy some further requirements. We now discuss several common criteria for assessing a theory's internal structure. Like our definition of a theory, they are inspired by the sciences, but apply to normative and evaluative theories as much as they apply to positive ones. Some of the criteria are so obvious that they are often left unacknowledged; nonetheless, it is useful to make them explicit.

Consistency: We require a good theory to be *logically consistent*. Formally, the set of statements constituting the theory must be capable of being simultaneously true. An obvious reason for requiring consistency is that anything follows from an inconsistent set of statements (*ex falso quodlibet*). Thus an inconsistent set, such as one containing both "*p*" and "*not p*", is of little use, whether for explanatory, predictive, evaluative, or prescriptive purposes. By entailing everything, it is too indiscriminate.[10]

Deductive closure: We require a good theory to be *deductively closed*. This means that any statement that is logically entailed by the theory also belongs to the theory.[11] The idea underlying deductive closure is that we want to be able to identify a theory with everything to which the theory is logically committed. If the theory asserts "*p*" and "*if p then q*", for example, then it should also assert "*q*". We would consider a theory defective if it were committed to the first two statements, but not to the third. The way we characterized Newton's and Rawls's theories had deductive closure built into it, since we characterized each as the set of all statements that are logically entailed by the relevant principles. Deductive closure is easy to achieve even when a theory is initially given in a form that violates it: we can re-define the theory as the set of all statements entailed by the original, non-deductively-closed formulation.[12]

Axiomatizability: This requirement is implicit in the final clause of our definition of a theory, which says that a theory should ideally be representable as the set of all implications of some underlying principles. Formally, a theory is *axiomatizable* if there exists a finite set of principles such that the entire theory can be expressed as their body of implications.[13] It should be evident from our earlier discussion that Newton's and Rawls's theories are usually presented in axiomatized form, as the bodies of implications of Newton's and Rawls's principles, respectively, perhaps together with some auxiliary assumptions. An

[10] Sometimes we may be prepared to lower the bar of consistency, by admitting some "contained" or "local" inconsistencies, as in so-called "paraconsistent" logics. Even then, we usually impose some weakened variant of a consistency requirement, to rule out those inconsistencies that are too global to preserve a theory's usefulness.

[11] A set of statements is *deductively closed* if it contains all its implications.

[12] The deductive-closure requirement also highlights, once more, why consistency matters. Since an inconsistent set of statements *entails* everything, deductive closure would force such a theory to *consist* of everything that can be expressed in the relevant language, which would amount to a completely uninformative theory. Again, if we were to use a paraconsistent logic, we might weaken the requirement of deductive closure but still retain some less demanding requirement in a similar spirit.

[13] See also Quine (1975). Note that any axiomatizable theory is deductively closed. Note, further, that logicians sometimes replace the "finiteness" requirement with a weaker "formal decidability" requirement; we set these technicalities aside.

axiomatizable theory can be presented in an informative manner, simply by specifying the set of principles from which it can be derived. If the theory could only be presented by brute enumeration of all its implications—typically infinitely many—there would be no succinct way of summarizing its content. A good theory illuminates its subject matter by giving us a manageable set of principles—a manageable theory formulation—that encodes the theory's entire propositional content.

Parsimony: We require a good theory to avoid any unnecessary complexity, and to be as simple as possible, in an appropriate sense of simplicity (on simplicity, see e.g. Baker 2013).[14] Scientists commonly care about parsimony, often under the label "Occam's razor", and Rawls, for instance, also emphasizes simplicity as a virtue of a normative theory.[15] What counts as "simple" may be different from context to context and may also depend on what the theory is about. We usually want to find the simplest theory able to account for its subject matter. To be illuminating, the theory ought to be simpler, for example, than the target phenomenon it seeks to account for, as well as simpler than its rival theories. A scientific theory, for instance, should be simpler and more succinct than an enumeration of all the empirical facts it seeks to explain; otherwise it cannot play any explanatory role. Likewise, a good normative theory should be simpler than an enumeration of all case-specific normative judgements. The relevant bar of simplicity may be adjusted depending on the theory's subject matter.

The present list of criteria for the internal assessment of theories is only illustrative, not exhaustive, but given space constraints, we now move on to external criteria.

5.2 External Criteria

We defined a theory as a set of statements that is a candidate for playing some theoretical or practical role. Implicit in this definition is the idea that there is something the theory is about: any theory is intended to represent, summarize, or capture something "outside the theory". It may capture this correctly, in which case the theory is true, correct, or externally valid, or it may fail to do so, in which case it is false, incorrect, or externally invalid.

What exactly a given theory is intended to represent needs to be spelt out further. In the case of a physical theory, the answer is relatively straightforward, especially if we accept "scientific realism": it is intended to represent certain physical facts about the world, such as facts about how physical objects behave in response to each other (see e.g. Chakravartty 2013). In the case of a normative or evaluative theory, the picture is more complicated. If we are realists about normative or evaluative matters, we may say that the theory is intended to represent some theory-independent moral facts. If we are not realists about normative or evaluative matters, it is harder to specify what a normative or evaluative theory is intended

[14] Although axiomatizability of a theory in terms of some easily expressible principles is one of the marks of parsimony, axiomatizability is not a sufficient condition for parsimony.

[15] In his discussion of why his principles of justice are preferable to utilitarian principles, Rawls says that "reasonable risk aversion may be so great, once the enormous hazards of the decision in the original position are fully appreciated, that the utilitarian weighting may be, for practical purposes, so close to the difference principle as to make the *simplicity* of the latter … decisive in its favour" (Rawls 1999, 144 emphasis added).

to represent (on moral realism and anti-realism see, respectively, Sayre-McCord 2011; Joyce 2009).

Yet the very idea of a theory breaks down unless we assume that there is something potentially representable by it, however observer-dependent or socially constructed it might be.[16] If we were nihilists, to take an extreme example of the denial of any normative or evaluative facts, we would not be able to engage in normative or evaluative theorizing in earnest.

Thus, in this section, we assume that normative or evaluative theories are *truth-apt*: it makes sense to ask whether they are true—or, some might prefer to say: correct or externally valid. We thus accept a form of "cognitivism" about such theories. This assumption is still compatible with a variety of views about the "meta-ethical" status of normative or evaluative judgements. We return to some of these issues in section 7.1. Granting, then, that there is some standard of correctness by which we can assess normative or evaluative theories (an ontological assumption), we still need to know how to do this assessment (an epistemological question). We now review several methods of testing a theory for external validity.

Taking intuitive judgments as strict evidence: According to this method, the test for a normative or evaluative theory is whether it fits our intuitive judgements about the relevant normative or evaluative matters. On this approach, our intuitive judgements have the same status as empirical observations in science. In science, a theory is *empirically adequate* if it entails the correct observation statements (see e.g. Quine 1975; van Fraassen 1980). Similarly, in moral and political theory, we might call a theory *normatively* or *evaluatively adequate* if it entails the correct normative or evaluative statements. According to the strict-evidence method, these are precisely the normative or evaluative statements supported by our intuitive judgements (cf. the discussion in Dworkin 1975). Although simple and analogous to familiar scientific methods, this method has some problems. First, while we may be confident in some of our normative or evaluative judgements, other judgements may be more tentative, and in some cases—especially when the issue is less familiar—we may not have any firm intuitions at all. Second, our intuitive judgements may be subject to biases and framing effects, which may cast further doubt on their reliability. Third, our normative or evaluative judgements may not be consistent with one another, or they may entail other judgements that we reject on reflection; in such cases, the strict-evidence method provides no guidance at all.

Reflective equilibrium: A more refined method of theory testing is the reflective-equilibrium method. It does not treat our intuitions as independent evidence, prior to the theory in question, but requires us to reach a "mutual fit" between the theory and our considered judgements. This works as follows. We begin with some initial theory, perhaps inspired by our initial intuitive judgements or given by some prima facie principles, then consider the implications of the theory and ask whether they are also in line with our judgements. If those implications fit our judgements, the process stops. It is more likely, however, that only

[16] A constructivist might take a theory to represent certain constructed facts. This is consistent even with the view that the theory itself is the "vehicle" by which those facts are being constructed. To develop that view further, one might draw, for instance, on parallels with Searle's analysis of declarative speech acts (which—roughly—bring certain facts into existence by representing them). See Searle (1995).

some of the theory's implications fit our judgements, while others do not. We then reassess both the theory and our judgements. In some cases, we may decide, on reflection, to revise the theory by changing some of the constituent principles, so as to bring the theory in line with the judgements we are unwilling to give up. In other cases, we may decide to overrule our judgements and embrace the theory's implications as our new considered judgements. A *reflective equilibrium* is reached when the implications of our possibly revised theory are in line with our possibly revised judgements. At least since Rawls's *Theory of Justice*, reflective equilibrium has been one of the most widely used methods in political theory (Rawls 1999, 15–18, 40–6; Daniels 2013). Its details can be spelt out in a variety of ways. For example, we may choose the units of assessment more narrowly or more broadly. We can either search for a reflective equilibrium involving a narrowly specified theory, constituted by a small number of principles together with very few auxiliary assumptions, or we can search for a reflective equilibrium involving a more broadly specified theory, constituted by a larger number of principles and further additional assumptions—which in turn may be interpreted as a conjunction of multiple theories, covering a wider domain of issues. Political theorists sometimes speak of "narrow" reflective equilibrium in the first case, and "wide" reflective equilibrium in the second (Daniels 2013). Similarly, we may take different views on which kinds of judgements—especially whose judgements—should serve as input to this method; more on this shortly. Although the reflective-equilibrium method is consistent with the idea that we arrive at our normative or evaluative theories through careful deliberation, we may be worried about the possible arbitrariness of its outcome, since there may not always be a unique equilibrium. In some cases, we may not be able to reach any equilibrium at all (the *non-existence* problem), such as when we theorize about genuine moral dilemmas and vacillate between different theories that each fit only some of our judgements while conflicting with others. In other cases, there may exist more than one equilibrium, in that we can arrive at different "packages" of revised theories and judgements that each have the required "mutual fit" (the *non-uniqueness* problem). Arbitrary factors such as framing effects or the order in which we consider different implications of the theory may then affect which reflective equilibrium we end up with (the *path-dependence* problem).

Thought experiments and intuition pumps: Whether we opt for the strict-evidence method or the reflective-equilibrium method, we may sometimes wish to sharpen or clarify our intuitions or judgements. Thought experiments and real-world cases can serve as useful "intuition pumps" (Dennett 2013; Brownlee and Stemplowska forthcoming). Here, we consider some hypothetical or actual scenario that prompts strong normative or evaluative judgements. In the much-discussed "trolley problems", for example, we are asked to judge what actions, if any, would be permitted to prevent a run-away trolley from crashing into, and killing, a larger group of people, at the expense of leading it to crash into, and kill, a smaller group (Thomson 1985). We then use these judgements to test our relevant normative or evaluative theories, following either the strict-evidence method or the reflective-equilibrium method. The usefulness of intuition pumps, especially ones involving highly idealized, counterfactual scenarios has recently been the object of considerable controversy in political theory (Elster 2011). We return to this issue in section 6.

The relevant judgements: Both the strict-evidence method and the reflective-equilibrium method raise the question of which kinds of judgements, and whose judgements, to use in

testing our theories. Should we test our theories on the basis of relatively spontaneous judgements or on the basis of suitably "filtered" judgements, and how should that filtering take place (Rawls 1999, 42)? And should we use the political theorist's judgements (which might be affected by his or her ideological views) or society's (which might similarly be affected by biases), and in the latter case, which society should we focus on (Miller 1992; Walzer 1983)? For example, while the Rawls of *A Theory of Justice* arguably followed the former approach (relying on the political theorist's judgements), the Rawls of *Political Liberalism* subscribed to the latter (looking at society) (Rawls 1971/1999; Rawls 1996). Specifically, the later Rawls re-interprets his theory of justice as an articulation of the ideas implicit in the public culture of liberal democratic societies. The building blocks of his account of justice are explicitly "drawn" from, and supposed to be widely acknowledged within, the society for which that account is designed. Relatedly, political theorists disagree about whether the judgements to which they appeal in theory testing should "fit" the particular practice the theory is meant to regulate. If the answer to this question is positive, then the exercise of theory construction is best seen as an attempt to offer what Ronald Dworkin calls a "constructive interpretation" of existing political practices (Dworkin 1986, chap. 2; James 2005; Sangiovanni 2008). If the answer is negative, then the exercise of theory construction is best understood as an attempt to "discover" particular normative and evaluative truths, independently of existing social practices (on interpretation versus invention, see Walzer 1987). To illustrate the difference between the two approaches—at least on one reading of what sets them apart—consider the following judgement, made explicitly in relation to the practice of camping with friends: "we should institute a regime of shared ownership" (see G. A. Cohen 2009). On a practice-dependent/interpretive approach, this judgement should count as relevant evidence only in the construction of a "normative theory of camping", not in the construction of normative theories of other practices, such as socio-political relations within the state. For proponents of practice-independence, by contrast, all normative or evaluative judgements, including the one in question, have cross-contextual validity in the identification of what justice or other moral values demand (this point is made in Ronzoni 2012; see also Miller 2002).[17]

The applied-moral-philosophy method: A final method of justifying a theory in political theory is to show that it can be derived from some independently accepted moral principles or theory. A committed utilitarian or Kantian, for example, may regard a normative or evaluative theory as justified if and only if it can be derived from utilitarian or Kantian principles, which are treated as independently given. In this vein, Nozick says: "Moral philosophy sets the background for, and boundaries of, political philosophy. What persons may and may not do to one another limits what they may do through the apparatus of the state, or do to establish such an apparatus" (Nozick 1974, 6; see also Otsuka 2003, 3). Those theorists who view political theory as a subfield of moral philosophy will find this method appropriate. (Works in which political theory is conducted—at least to some extent—as applied moral philosophy include Singer 1972; G. A. Cohen 2008; Fabre 2012.) By contrast, if we consider the activity of political theory

[17] The distinction between practice-dependence and practice-independence is complex and much debated. Due to space constraints, we are unable to explore this complexity here.

to take place against the background of (reasonable) pluralism about moral matters, the applied-moral-philosophy method is problematic, since it relies on the acceptance of a specific moral theory (for a critique of the "applied moral philosophy" approach, see Williams 2005; see also Galston 2010).

6. Abstraction and Idealization in Political Theory

In political theory, as well as in other disciplines, theories are often abstract and/or idealized in certain respects. In this section, we explain what this means and discuss some methodological issues raised by abstraction and idealization.

6.1 Defining Abstraction and Idealization

Broadly following Onora O'Neill (1996, chap. 2), we say that a theory is *abstract* with respect to an issue—represented by a set of statements—if it is silent on that issue; formally, it has no implications at all for the given statements, implying neither any of these statements nor any of their negations. Newton's theory of physics, for example, is abstract with respect to the colours of the physical bodies whose motion it represents. A theory is *idealized* with respect to an issue—again represented by a set of statements—if it entails some (simplifying or limiting) falsehood about that issue; formally, it has a false implication for some of the given statements, implying one or more false statements among them or the negations of one or more true statements. A simple Newtonian theory of mechanics, for instance, may be idealized with respect to friction, entailing the absence of friction in the physical systems it represents, although friction is present in the real world.

Since theories are meant to simplify the world, most political theorists agree that abstraction is virtually unavoidable in theory construction, and an innocuous intellectual exercise (O'Neill 1996, chap. 2; see also the discussion in Stemplowska 2008). Idealization, unlike abstraction, is looked at with greater suspicion in political theory and is considered potentially problematic (O'Neill 1996, chap. 2).

6.2 The Worry about Idealization

The "danger of idealization" in political theory has been discussed primarily in the debate on "ideal versus non-ideal theory" (for an overview, see Valentini 2012). The debate is largely animated by the worry that resort to simplifying assumptions and idealized thought experiments or intuition pumps—which are common in contemporary political theory—will adversely affect the validity of the ensuing theories (see e.g. O'Neill 1996, chap. 2; Farrelly 2007; Mills 2005). For instance, John Rawls develops his theory of justice assuming full compliance with the resulting principles. This

makes critics wonder whether we can trust Rawls's principles to deliver correct and action-guiding prescriptions for the real world, where many people fail to abide by the demands of justice.

In assessing this worry, two points are worth making. First, like abstraction, some degree of idealization in our theories—in the form of simplifying assumptions—may play an important, and justified, heuristic role. Again, a simple newtonian theory of physics does not appear to be significantly undermined by its assuming friction away, especially to the extent that information about friction can potentially be re-introduced in refinements or applications of the theory. This suggests that, rather than reject idealizations as problematic from the outset, we must ask whether a theory contains the "right" idealizations, given its purpose (Robeyns 2008; Valentini 2009).

Second, to answer the question of which idealizations are "right" and which are not, we need to distinguish between three possible *loci*—or levels—at which idealizations can occur: (1) the theory itself, (2) the conditions of application of the theory's prescriptions (where those prescriptions are of the form "if such-and-such conditions hold, then such-and-such follows"), (3) the justification of the theory. Crucially, idealization at any one of these levels need not entail idealization at any of the others.

6.3 Rawls's Theory as an Example

Rawls's theory of justice, as described earlier, consists of the "equal liberty" (X), "fair equality of opportunity" (Y), and "difference" (Z) principles and their implications. The theory itself would be idealized—a level (1) idealization—*if and only if* these principles entailed false statements about their subject matter—for instance, if the statement "we ought to rearrange the tax system so as to benefit the worst off as much as possible" were false. Note, however, that the often-criticized "idealized" assumption of full compliance is not made within the theory itself, but occurs as part of the *justification* that Rawls offers for the theory. It is one of the assumptions made by the parties in the original-position thought experiment. So, it is an idealization at level (3). Rawls's principles, which generate his theory, do not imply any false claims about full compliance; hence we have no idealization at level (1) here. Nor do the prescriptions following from these principles presuppose full compliance for their *applicability*; hence we have no idealization at level (2) either (see Simmons 2010, 9–10 for discussion).

Similarly, consider Rawls's assumption that society exists under favourable historical and social conditions. Relative to existing war-torn, or desperately poor countries, this assumption is clearly false: it is an idealization. Rawls says explicitly that his principles of justice may not apply to societies in which the relevant favourable conditions are absent (Rawls 1999, 216). Does this make Rawls's *theory* idealized—an idealization at level (1)—and problematically so? Arguably, it does not, because the favourable-conditions idealization operates at the level of the *conditions of application* of the theory's prescriptions, that is, at level (2). The prescription that is entailed by the theory (under a careful formulation)—namely "*if favourable conditions hold*, justice demands X, Y, and Z"—is still true. Although the conditional nature of this prescription limits

the scope of application of the theory, it does not make the theory itself idealized, by generating *false* prescriptions.

6.4 The Worry Reassessed

We suspect that, although worries about idealization in political theory are frequently expressed as complaints about "*theories* being idealized", they actually tend to target idealizations at levels (2) and (3), rather than (1)—that is, at the levels of the theories' conditions of application and the justifications offered for those theories, not at the level of the theories themselves.

Of course, idealizations at levels (2) and (3) often make the target-theories somewhat irrelevant to the real world, by rendering them insufficiently action-guiding in real-world circumstances. While this lack of guidance may be a genuine shortcoming, it is not accurately captured by the claim that the theories themselves are idealized. Rather, in the case of a level (2) idealization, the theories entail true prescriptions of an "if-then" sort, whose antecedent conditions—the "if" clauses—do not hold in real-world circumstances. And in the case of a level (3) idealization, the theories may be insufficiently justified, in that the justifications offered for them—such as highly contrived thought experiments—are too idealized to allow inferences for the real world.

Our discussion highlights the importance of clarity about the role that idealizations play in the defence and formulation of one's theory. Consider a theory of justice prescribing "*p*", defended on the assumption that there is no reasonable disagreement about justice within society, and yet formulated in universal terms: "justice *always* demands that *p*". This theory does indeed run the risk of being problematically idealized, if it turns out that the presence of reasonable disagreement makes a morally relevant difference to what justice demands. The theory, in that case, would have false implications in a number of situations, namely those involving reasonable disagreement. A true principle, by contrast, would only say: "if there is no reasonable disagreement, justice demands that *p*".

7. The Significance of Disagreement in Political Theory

We noted earlier that, while political theory is sometimes viewed as a subfield of moral philosophy, another view is that political theory is distinct from moral philosophy in that the conditions of theorizing are different (see e.g. Williams 2005; Larmore 2013). On this view, political theory, unlike—or much more than—moral philosophy, is conducted against the background of (reasonable) pluralism in society. Disagreement about normative and evaluative matters, it is said, has a different status in political theory than in (mainstream) moral philosophy. The nature of disagreement in society constrains what normative or evaluative principles can be defended in political theory. In this section, we explore this view and discuss its methodological implications.

7.1 Disagreement: "Constitutive of Assertability Conditions" Rather Than "Epistemic"

While moral philosophers seek to answer questions such as "what ought we to do in a given situation", political theorists are faced with questions such as "what ought we do, given that we do not agree about what we ought to do" (see e.g. Waldron 1999; Larmore 2013). At first sight, we might be perplexed by this alleged difference between political theory and moral philosophy. Are there not equally big disagreements in moral philosophy? Practically every well-known moral dilemma has the property that different people take different views on how it should be resolved, and the question of what one should do in such cases of disagreement arises also in the moral realm.

Yet, the difference between moral philosophy and political theory may be said to be the following. In moral philosophy, we commonly (though not universally) make the assumption that, among the many different rival normative or evaluative theories, one is the independently correct or true theory. The task for the moral philosopher is to identify that theory. *If* we conduct moral philosophy on this assumption, disagreement is of a "merely" epistemic kind. There is a fact about what the right answer to any normative or evaluative question is; we may just have different beliefs about that fact.

It is less clear—so the argument goes—whether the same assumption can be made in political theory. On this picture, disagreement in political theory may be viewed, not merely as "epistemic", reflecting different beliefs about the same truth, but as partly "constitutive" of the correctness conditions—or warranted assertability conditions—of normative claims themselves. On this view, whether a normative or evaluative theory in political theory is correct (or assertable) depends, in part, on the society in which the theory is to be applied, and specifically on the level of reasonable disagreement in that society; we define the notion of *reasonableness* in section 7.3.

A key desideratum here is that our theories be, at least in principle, acceptable to individuals holding reasonable but conflicting moral views within pluralistic societies. What counts as a correct normative or evaluative principle in a society with little reasonable disagreement need not always count as correct in a highly pluralistic society, where the range of reasonable disagreement is greater. In light of this, some theorists avoid using the notion of truth in political theory altogether, and prefer to replace it with other, less loaded notions, like reasonableness or reasonable acceptability (Rawls 1996; cf. J. Cohen 2009).

But what might justify the shift from the "epistemic" to the "constitutive" status of disagreement in the political domain? At least two answers are possible. First, the shift might be *morally justified*. On this view, a commitment to respect for persons places a burden of "reasonable acceptability" on the principles put forward by the political theorist, insofar as these principles may be *permissibly enforced* against individuals, for instance through state action. The idea is that political theory focuses on the development of "enforceable rules", and these rules are normatively appropriate (meet the relevant criterion of correctness) only if they are acceptable to reasonable individuals to whom they apply.

Second, reasonable disagreement may be taken to constrain correctness or warranted assertability in political theory for *pragmatic reasons*. After all, most of the principles put forward by political theorists are meant to help regulate social life in complex and highly pluralistic societies. A normative or evaluative theory that appeals only to people of a

particular moral persuasion would be of little use in this respect; it could not give rise to a durable and stable social order except through autocratic imposition. And so again, acceptability to individuals with competing reasonable views may be deemed a criterion of correctness of the theory.

The inspiration for many of the foregoing reflections can be found in John Rawls's second major book, *Political Liberalism* (1993/1996). We now offer a brief discussion of Rawls's treatment of the relevance of pluralism to political theorizing.

7.2 The Distinction Between "Political" and "Comprehensive" Theories

One of the key innovations of *Political Liberalism*, compared to *A Theory of Justice*, is Rawls's insistence on his theory being "political not metaphysical" (Rawls 1985). This means two things. First, the theory does not adjudicate all aspects of interpersonal conduct (as a "comprehensive" moral theory might do) but concerns only the "public" realm (most notably, the basic structure of society). Second, the theory is grounded *not* in any "comprehensive" moral theory ("doctrine"), on which there is likely to be deep disagreement in any pluralistic liberal society, but in ideals that are drawn from the public culture of such a society. Thus a "political" theory is typically (i) restricted to a smaller domain of issues and (ii) less morally and metaphysically "loaded" (thereby "thinner") than a "comprehensive" theory. In particular, it refrains from taking a stand on issues that are too controversial.

For example, a theory of justice that adjudicates all aspects of personal, and not just political and social, life would be "comprehensive", as would be a theory based on the metaphysically loaded premise that human beings are morally equal because they are created "in the image of God". Both theories would be the object of *reasonable disagreement* within a pluralistic society. By contrast, a theory of justice that focuses on the basic structure of society and is based on a commitment to "citizens' freedom and equality" would count as "political", since its domain does not "over-reach" and its premises are entrenched in the public culture of liberal democracies and arguably shared by individuals holding competing yet reasonable comprehensive doctrines. But what does the notion of "reasonableness" stand for?

7.3 The Notion of "Reasonable Disagreement"

The notion of "reasonableness" is hard to pin down in general, and this is true in Rawls's case as well (for discussion, see Gaus 1996, 131–132; Gaus 1999). In particular, the notion may be interpreted in epistemological and/or moral ways. Under an epistemological interpretation, something is *reasonable* if it is consistent with a proper use of reason in light of the evidence available. Rawls believes that, because of what he calls the "burdens of judgment", we should not expect agreement between the views of different people, even when these views are all developed by consistently applying the powers of reason (Rawls 1996, Lecture 2, sec. 2). Under a moral interpretation, *reasonableness* refers to a view's compatibility with certain fundamental normative requirements, such as respect for citizens as free and equal, or a

commitment to mutual justification (Rawls 1996, Lecture 2, sec. 3). This would make views that reject such a commitment unreasonable.

Both interpretations of "reasonableness" have virtues and vices, especially in the context of what we earlier called the "moral" justification for deeming acceptability in the context of reasonable disagreement an assertability condition for theories within political theory. As far as the epistemological interpretation of reasonableness is concerned, on the positive side, it offers a compelling rationale for treating only *some* contested views as "worthy of respect", namely those with a given epistemic pedigree. On the negative side, this interpretation potentially allows morally repugnant views to count as reasonable, assuming they involve no breach of reason, purely epistemically understood. If a necessary condition for the overall acceptability of a normative or evaluative theory is that it is acceptable to all individuals with reasonable views, an acceptable theory must then potentially appeal even to individuals with morally repugnant views, and this may be too much to ask.

As far as the moral interpretation of reasonableness is concerned, on the positive side, it allows us to "filter out" morally repugnant views. But, on the negative side, it does so at the cost of referring to substantive values that themselves require justification. The suspicion is that, on a moralized interpretation, the "reasonable" simply corresponds to the moral commitments that the liberal theorist considers non-negotiable: it is ad hoc (see Mouffe 2005). Again, we do not take a stand on this, but note that these difficulties threaten the plausibility of defining the overall acceptability of a normative or evaluative theory simply in terms of its acceptability to all individuals with reasonable views; identifying those reasonable views may itself rely on moral premises.

7.4 The Quest for an "Overlapping Consensus"

Finally, turning to the "pragmatic" justification for taking disagreement to have a "constitutive" rather than purely "epistemic" status in political theory, we conclude with some remarks on Rawls's notion of an "overlapping consensus". An *overlapping consensus* on a normative or evaluative theory in a social or political domain occurs when this theory is endorsed from the perspective of different reasonable comprehensive doctrines held in society. A normative theory that cannot be endorsed from different such perspectives could not hope to gain sufficient support and to offer a stable basis for social organization in pluralistic liberal democracies (Rawls 1996, Lecture 4). The notion of an overlapping consensus operationalizes the idea that the criterion of correctness of a "political" theory is its acceptability to individuals who hold competing but reasonable comprehensive moral theories.

8. Concluding Remarks

We have reviewed the methodology of analytic political theory from what we hope is a somewhat novel and helpful angle. By drawing on ideas from the philosophy of science, we have attempted to highlight the ways in which theorizing in political theory relates to

theorizing in other areas of philosophy and positive science. We have also reviewed some recent debates and controversies within the political-theory literature, which has only recently given greater attention to methodological questions. Our hope is that this article will prove to be a clarifying contribution to the growing methodological debate in political theory.

References

Arrow, Kenneth Joseph. 1951. *Social Choice and Individual Values*. New York: John Wiley & Sons.

Baehr, Amy R. 2013. "Liberal Feminism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2013. <http://plato.stanford.edu/archives/win2013/entries/feminism-liberal/>. Accessed September 24, 2015.

Baker, Alan. 2013. "Simplicity." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2013. <http://plato.stanford.edu/archives/fall2013/entries/simplicity/>. Accessed September 24, 2015.

Blackorby, Charles, Walter Bossert, and David J. Donaldson. 2005. *Population Issues in Social Choice Theory, Welfare Economics, and Ethics*. New York: Cambridge University Press.

Bratman, Michael. 1999. *Faces of Intention: Selected Essays on Intention and Agency*. Cambridge: Cambridge University Press.

Brownlee, Kimberley, and Zofia Stemplowska. Forthcoming. "Trapped in an Experience Machine with a Famous Violinist: Thought Experiments in Normative Theory." In *Research Methods in Analytical Political Theory*, edited by Adrian Blau. Cambridge: Cambridge University Press.

Carter, Ian. 2012. "Positive and Negative Liberty." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2012. <http://plato.stanford.edu/archives/spr2012/entries/liberty-positive-negative/>. Accessed September 24, 2015.

Chakravartty, Anjan. 2013. "Scientific Realism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2013. <http://plato.stanford.edu/archives/sum2013/entries/scientific-realism/>. Accessed September 24, 2015.

Christiano, Thomas. 2008. "Democracy." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2008. <http://plato.stanford.edu/archives/fall2008/entries/democracy/>. Accessed September 24, 2015.

Cohen, G. A. 1988. *History, Labour, and Freedom: Themes from Marx*. Oxford: Clarendon Press.

Cohen, G. A. 1995. *Self-Ownership, Freedom, and Equality*. Cambridge: Cambridge University Press.

Cohen, G. A. 2008. *Rescuing Justice and Equality*. Cambridge, MA: Harvard University Press.

Cohen, G. A. 2009. *Why Not Socialism?* Princeton, N.J.: Princeton University Press.

Cohen, Joshua. 2009. "Truth and Public Reason." *Philosophy & Public Affairs* 37 (1): 2–42.

Daniels, Norman. 2013. "Reflective Equilibrium." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2013. <http://plato.stanford.edu/archives/win2013/entries/reflective-equilibrium/>. Accessed September 24, 2015.

Dennett, Daniel C. 2013. *Intuition Pumps And Other Tools for Thinking*. New York: W. W. Norton.

Dworkin, Ronald. 1975. "The Original Position." In *Reading Rawls: Critical Studies on Rawls's A Theory of Justice*, edited by Norman Daniels, 16–52. Stanford, CA: Stanford University Press.

Dworkin, Ronald. 1986. *Law's Empire*. Cambridge, MA: Harvard University Press.

Elster, Jakob. 2011. "How Outlandish Can Imaginary Cases Be?" *Journal of Applied Philosophy* 28 (3): 241–58.

Fabre, Cécile. 2012. *Cosmopolitan War*. Oxford: Oxford University Press.

Farrelly, Colin. 2007. "Justice in Ideal Theory: A Refutation." *Political Studies* 55 (4): 844–64.

Gallie, W. B. 1955. "Essentially Contested Concepts." *Proceedings of the Aristotelian Society* 56 (1): 167–98.

Galston, William A. 2010. "Realism in Political Theory." *European Journal of Political Theory* 9 (4): 385–411.

Gärdenfors, Peter. 2000. *Conceptual Spaces: The Geometry of Thought*. Cambridge, MA: MIT Press.

Gaus, Gerald F. 1996. *Justificatory Liberalism : An Essay on Epistemology and Political Theory*. New York: Oxford University Press.

Gaus, Gerald F. 1999. "Reasonable Pluralism and the Domain of the Political: How the Weaknesses of John Rawls's Political Liberalism Can Be Overcome by a Justificatory Liberalism." *Inquiry* 42 (2): 259–84.

Gerring, John. 1999. "What Makes a Concept Good? A Criterian Framework for Understanding Concept Formation in the Social Sciences." *Polity* 31 (3): 357–93.

Gilabert, Pablo, and Holly Lawford-Smith. 2012. "Political Feasibility: A Conceptual Exploration." *Political Studies* 60 (4): 809–25.

Gilbert, Margaret. 1989. *On Social Facts*. Princeton, NJ: Princeton University Press.

Goodin, Robert E. 2009. *The Oxford Handbook of Political Science*. Oxford: Oxford University Press.

Harsanyi, John C. 1955. "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility." *Journal of Political Economy* 63 (4): 309–21.

James, Aaron. 2005. "Constructing Justice for Existing Practice: Rawls and the Status Quo." *Philosophy & Public Affairs* 33 (3): 281–316.

Joyce, Richard. 2009. "Moral Anti-Realism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2009. <http://plato.stanford.edu/archives/sum2009/entries/moral-anti-realism/>. Accessed September 24, 2015.

Larmore, Charles. 2013. "What Is Political Philosophy?" *Journal of Moral Philosophy* 10 (3): 276–306.

Leopold, David, and Marc Stears. 2008. *Political Theory: Methods and Approaches*. Oxford: Oxford University Press.

Linz, Juan J., and Alfred Stepan. 1996. *Problems of Democratic Transition and Consolidation: Southern Europe, South America, and Post-Communist Europe*. Baltimore: Johns Hopkins University Press.

List, Christian. 2011. "The Logical Space of Democracy." *Philosophy & Public Affairs* 39 (3): 262–97.

List, Christian. 2013. "Social Choice Theory." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2013. <http://plato.stanford.edu/archives/win2013/entries/social-choice/>. Accessed September, 24, 2015.

List, Christian, and Philip Pettit. 2011. *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford: Oxford University Press.

Lovett, Frank. 2013. "Republicanism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2013. <http://plato.stanford.edu/archives/spr2013/entries/republicanism/>. Accessed September, 24, 2015.

MacCallum, Gerald C. 1967. "Negative and Positive Freedom." *The Philosophical Review* 76 (3): 312–34.

McDermott, Daniel. 2008. "Analytic Political Philosophy." In *Political Theory: Methods and Approaches*, edited by David Leopold and Marc Stears, 11–28. Oxford: Oxford University Press.

Margolis, Eric, and Stephen Laurence. 2012. "Concepts." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2012. <http://plato.stanford.edu/archives/fall2012/entries/concepts/>. Accessed September, 24, 2015.

Miller, David. 1992. "Distributive Justice: What the People Think." *Ethics* 102 (3): 555–93.

Miller, David. 2002. "Two Ways to Think About Justice." *Politics, Philosophy & Economics* 1 (1): 5–28.

Mills, Charles W. 2005. "'Ideal Theory' as Ideology." *Hypatia* 20 (3): 165–83.

Mouffe, Chantal. 2005. "The Limits of John Rawls's Pluralism." *Politics, Philosophy & Economics* 4 (2): 221–31.

Nozick, Robert. 1974. *Anarchy, State, and Utopia*. New York: Basic Books.

O'Neill, Onora. 1996. *Towards Justice and Virtue: A Constructive Account of Practical Reasoning*. Cambridge: Cambridge University Press.

Okasha, Samir. 2002. *Philosophy of Science: A Very Short Introduction*. Oxford: Oxford University Press.

Oppenheim, Felix E. 1981. *Political Concepts: A Reconstruction*. Oxford: Basil Blackwell.

Otsuka, Michael. 2003. *Libertarianism Without Inequality*. Oxford: Clarendon Press.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Clarendon Press.

Pettit, Philip. 1993. *The Common Mind: An Essay on Psychology, Society, and Politics*. New York: Oxford University Press.

Pettit, Philip. 1997. *Republicanism. A Theory of Freedom and Government*. Oxford: Clarendon Press.

Pettit, Philip. 2011. "The Instability of Freedom as Noninterference: The Case of Isaiah Berlin." *Ethics* 121 (4): 693–716.

Quine, W. V. 1975. "On Empirically Equivalent Systems of the World." *Erkenntnis* 9 (3): 313–28.

Rawls, John. 1985. "Justice as Fairness: Political Not Metaphysical." *Philosophy and Public Affairs* 14 (3): 223–51.

Rawls, John. 1996. *Political Liberalism*. New York: Columbia University Press.

Rawls, John. 1999. *A Theory of Justice*. Oxford: Oxford University Press.

Robeyns, Ingrid. 2008. "Ideal Theory in Theory and Practice." *Social Theory and Practice* 34 (3): 341–62.

Roemer, John E. 1998. *Equality of Opportunity*. Cambridge, MA: Harvard University Press.

Ronzoni, Miriam. 2012. "Life Is Not a Camping Trip—on the Desirability of Cohenite Socialism." *Politics, Philosophy & Economics* 11 (2): 171–85.

Sangiovanni, Andrea. 2008. "Justice and the Priority of Politics to Morality." *Journal of Political Philosophy* 16 (2): 137–64.

Sayre-McCord, Geoff. 2011. "Moral Realism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2011. <http://plato.stanford.edu/archives/sum2011/entries/moral-realism/>. Accessed 24 September, 2015.

Searle, John R. 1995. *The Construction of Social Reality*. New York: Free Press.

Sen, Amartya K. 1970. *Collective Choice and Social Welfare*. San Francisco: Holden-Day.
Sen, Amartya K. 1980. "Equality of What?" In *The Tanner Lectures on Human Values, Volume I*, edited by Sterling M. McMurrin, 195–220. Cambridge: Cambridge University Press.
Sen, Amartya K. 1987. *The Standard of Living*. Edited by Geoffrey Hawthorn. Cambridge: Cambridge University Press.
Simmons, A. John. 2010. "Ideal and Nonideal Theory." *Philosophy & Public Affairs* 38 (1): 5–36.
Singer, Peter. 1972. "Famine, Affluence, and Morality." *Philosophy & Public Affairs* 1 (3): 229–43.
Southwood, Nicholas. 2013. "Democracy as a Modally Demanding Value." *Noûs* Online early: n/a–n/a.
Stemplowska, Zofia. 2008. "What's Ideal About Ideal Theory?" *Social Theory and Practice* 34 (3): 319–40.
Thomson, Judith Jarvis. 1985. "The Trolley Problem." *The Yale Law Journal* 94 (6): 1395–415.
Tollefsen, Deborah Perron. 2002. "Collective Intentionality and the Social Sciences." *Philosophy of the Social Sciences* 32 (1): 25–50.
Tuomela, Raimo. 2007. *The Philosophy of Sociality: The Shared Point of View*. New York: Oxford University Press.
Valentini, Laura. 2009. "On the Apparent Paradox of Ideal Theory." *Journal of Political Philosophy* 17 (3): 332–55.
Valentini, Laura. 2012. "Ideal vs. Non-Ideal Theory: A Conceptual Map." *Philosophy Compass* 7 (9): 654–64.
Van Fraassen, Bas C. 1980. *The Scientific Image*. Oxford: Clarendon Press.
Waldron, Jeremy. 1999. *Law and Disagreement*. Oxford: Clarendon Press.
Walzer, Michael. 1983. *Spheres of Justice: A Defense of Pluralism and Equality*. New York: Basic Books.
Walzer, Michael. 1987. *Interpretation and Social Criticism*. Cambridge, MA: Harvard University Press.
Williams, Bernard Arthur Owen. 2005. *In the Beginning Was the Deed: Realism and Moralism in Political Argument*. Princeton, N.J.: Princeton University Press.

CHAPTER 28

PHILOSOPHY AND PSYCHOLOGY

LOUISE ANTONY AND GEORGES REY

1. Introduction

This chapter will be concerned with the relationship between philosophy and psychology, especially with the increasingly interesting ways in which psychological research seriously affects philosophical speculation, confirming it in some cases, posing serious problems for it in others. Limitations of time, space, and expertise require that our discussion be confined largely to the relation of experimental psychology to the tradition of so-called "analytic philosophy" that has dominated the field in anglophone countries for the last century.[1]

One reason for the close connection between philosophy and psychology is that both disciplines have been equally concerned with fundamental methodological questions: for example, Just what is the mind? Should it be understood through a priori reflection? What are the (different?) roles of introspective reports, observations of behavior, and neurophysiological data? We'll therefore begin, in §2, with a brief history of the co-evolution of philosophical and empirical theories of psychology: the introspectionist beginnings (§2.1), and then the various forms of behaviorism that dominated the mid-twentieth century (§§2.2; 2.3; 2.4), up through the work of Noam Chomsky and his followers (§2.5), and some of its methodological consequences. In §3, we'll look at the "computational" and "functionalist" approaches to the mind that grew out of both Chomsky's work and the development of the computer, turning in §4 to some of the philosophically important research of the resulting "cognitive science" regarding the character of the computations and the architecture of the mind (§4.1), and relations to epistemology (§4.2), biology (§4.3), and moral psychology (§4.4).

Another way in which psychology has influenced philosophical methodology is by uncovering empirical phenomena that raise difficulties for traditional philosophical

[1] For discussion of work in other traditions, see other entries, e.g. "Phenomenonology," (Chapter 10).

conceptions of the mind. In §5, we will discuss two important cases, experimental results bearing on the nature of consciousness and its supposed unity (§5.1), and phenomena of "self-blindness" and "blindsight," both of which raise surprising empirical questions about our introspective access to our mental states (§5.2).

2. Brief History

2.1 Introspection and Its Limits

Philosophy and psychology have been intertwined from the start. Speculations about the nature and existence of the soul have their roots in religious traditions, and certainly, by the time of Plato and Aristotle, in substantive theories about the "divisions of the soul," and about the objects of thought as well as about its origins, in innate "recollection" and/or in some kind of "enformed" relation with the properties of the world. These speculations continued through the Middle Ages, notably with Ockham and Aquinas, and then, beginning in the seventeenth century, with the familiar debate between the "rationalists" and "empiricists," as well as with the more diverse and complex proposals of, for example, Kant, Hegel, Schopenhauer, and Nietzsche in the eighteenth and nineteenth. At least by the end of the nineteenth century, there were departments of "philosophy and psychology," and William James, a professor in such a department at Harvard, made contributions to both these fields (psychology became a separate department at Harvard only in 1936).

A striking fact about these traditional hypotheses about the mind is how few of them were based on any careful experimentation of the sort that has been taken for granted for the past five hundred years in the "natural sciences" such as physics, chemistry, and biology. This was partly due to the fact that many philosophers regard(ed) their speculations as *a priori*: consideration of, for example, conditions necessary for the possibility of experience seemed to many of them not to depend upon the empirical character of this or that actual experience. But it was also due in part to a widespread presumption that mental states and processes were somehow immediately available, "transparent," at least to an introspective or reflective first-person point of view (these views are both challenged and defended in the present day; see §5.2).

In any case, it was no accident that, when nineteenth-century scientists, such as Fechner (1860/1965), Wundt (1892/1901) and Titchener (1908/1973) tried to apply the empirical methods of the natural sciences, they turned to experimentally controlled introspection. "Introspectionism," as their movement was called, did achieve some important findings, notably in "psychophysics," or the study of the relation of perceptual experiences to real stimuli. The most famous results, now known as the "Fechner/Stevens Laws," logarithmically related perceived to actual magnitudes. And many have found insightful the introspective discussions in William James' (1890/1981) famous *Principles of Psychology*. But it soon emerged that introspective reports of other processes, such as thought, imagery, reasoning, or emotion, were far too unreliable to afford a basis for any serious science.[2]

[2] See Schwitzgebel, 2010, for a rich, extended review of the topic.

There were two other essentially introspectivist approaches to psychology, roughly speaking, the "Hermeneutic" tradition that emerged from the work of Edward Titchener (1909), Wilhelm Dilthey (1927/76) and Max Weber (1904/49), and then the influential tradition of Freudian psychoanalysis. Hermenueticists argued that there was a fundamental difference between the "explanatory" ("erklärung") method of the natural sciences and the "empathic" ("verstehen" or "Hermeneutic") method of the social sciences. In fields such as history and psychology, they claimed that we come to understand a person's actions, not by subsuming them under the kinds of general laws discovered in the natural sciences, but by putting ourselves in agents' shoes, and "interpreting" the world as they did (see the chapters on "Phenomenology" and "History of Ideas: a Defense"). Thus, for example, even if it is turns out to be a stable, lawlike generalization that suicide rates are greater in time of peace than in times of war (as Durkheim, 1897/1951, observed), many would think we achieve a better understanding of it by coming to understand the thoughts and motives that lead people to it, as one finds in narrative histories. In analytic philosophy, aspects of the tradition surface in, for example, Davidson (1963, 2001), Dennett (1987), Rorty (1991), and McDowell (1994).

Freud's (1916/89) work and the clinical practice of psychoanalysis might be regarded as a kind of extension of both the introspectivist and Hermeneutic approaches, since theories in this domain are largely based on empathic interpretations of clinical data and the patient's ultimate acceptance of them.[3] Psychoanalysis has had significant effects on large parts of European philosophy, particularly, for example, in the work of Lacan, Derrida and Foucault. Foucault (1961/2006), for example, famously challenged the idea of individual mental "sickness" and put forward a socio-political analysis of mental "deviance," explaining the institutionalization of the "mentally ill" in terms of power relations. And Ian Hacking (1998) also has argued that pathologies of the mind are to an important extent social constructions (not that this makes them "unreal"). The entire project of classifying mental afflictions has engaged many philosophers of science, interested in the question whether the taxa of clinical psychology are simply behavioral syndromes, or are the correspondents of some "biologically real" pathology. Dominic Murphy (2006) has recently published a careful theoretical treatment of these methodological questions, and has argued that clinicians are indeed aiming at a principled "nosology" of mental afflictions.

Despite the enormous influence of psychoanalysis in anglophone culture in the twentieth century, however, the reliance on uncontrolled and often insufficiently documented clinical material has made many analytic philosophers and experimental psychologists wary of the conclusions (see Grünbaum, 1984, and Erwin, 1996 for discussion; but see also M. Levine, 2002, for a representative sample of recent analytic discussions of Freudian ideas, and Stern, 1985/1998, for an effort to integrate psychoanalytic issues into the experimental work on infant cognition that we will mention in §4.4).

[3] Freud hoped for some neurological vindication of his views, but realized this was well beyond the scientific methodology of his day.

2.2 Scientific Behaviorism

Partly in response to the failure of introspectionism, and to dissatisfaction with the other kindred approaches, psychologists such as Edward Thorndike (1911) and John Watson (1913) argued that a serious science of psychology should not be based upon "private" mental events, accessible only to the subject of them. Instead, they sought to describe regularities between an organism's publicly observable behavior and the environmental variables upon which it depends. In the hands of B. F. Skinner (1938) the approach came to called "Scientific" (or "Radical") Behaviorism," an approach that resonated with the then prevailing Positivistic skepticism about unobservables.

The overarching theoretical principle of the approach was Thorndike's "Law of Effect," which said, roughly, that behavior that was followed by positive consequences for the organism was likely to be repeated. Thus, a pigeon's tendency to peck at a lever whenever a light was flashed could be explained in terms of the patterns of rewards the pigeon had received for displaying that behavior in the past. Some stimuli, like food, were natural or "unconditioned" reinforcers. Initially neutral stimuli, like the sound of a bell, could themselves become "secondary" reinforcers through "conditioning," the repeated pairing of a neutral stimulus with an unconditioned reinforcer," and then, by the iterated pairing of this with still further stimuli, behavior could be "shaped," and "chains" of behavioral responses could accrete, giving rise to behavior that could seem quite remote from the original unconditioned reinforcement: a pigeon could progress from pecking at a bar to playing ping pong.[4]

Behaviorists held that shaping and chaining could provide a full account of all the complex behaviors we observe in humans and animals. The theory was attractive to the behaviorists, not only because it afforded a basis for an objective psychology, but also because reinforcement theory seemed to capture the traditional empiricist view, widely assumed to be a truism at the time, that knowledge was based on experience, a view that came to be seriously challenged, as we'll see shortly.

If this scientific program had been successful, it would have had profound implications for both philosophy and for standard folk psychological explanations (we'll return to its problems in §§2.5 and 2.6). Indeed, the presumption that something like it was true influenced many philosophers, notably Wittgenstein (1953), who often appealed to "training" as an explanation of human capacities (see, e.g. 1953:§§5–9, 86, 630), and Ryle (1949), who claimed "we find something implausible in the promise of hidden discoveries yet to be made of the hidden causes of our action" (1949:325). Unlike Skinner, however, Wittgenstein and Ryle did not take themselves to be advancing scientific claims. They were concerned instead to diagnosis what they regarded as traditional philosopher's misunderstanding of ordinary mental talk, which they proposed to understand and sometimes "analyze" in terms of behavior.

2.3 Analytical Behaviorism and Wittgenstein

Scientific Behaviorism was a specific scientific theory in psychology. Analytical Behaviorism was a logically independent *semantic* thesis about the meanings of mental

[4] One can see a display of this at YouTube video: <https://www.youtube.com/watchv=vGazyH6fQQ4> (accessed October 25, 2015).

terms, and it had its source in what has been a perennial idea since early empiricism, the *verificationist theory of meaning*, whereby the meaning of a sentence is regarded as consisting of the conditions that confirm or disconfirm it. This view went hand in glove with a skepticism about positing "unobservable" entities, both in science and in ordinary thought (see chapter 3 entitled "Methodology in Nineteenth- and Early Twentieth-Century Analytic Philosophy"). Thus, the meanings of claims about minds were thought to consist in what seemed to be the only objective evidence we could have of claims for them, namely, overt, physically characterized behavior.

The proposed analyses of mental claims never succeeded, for two reasons. First, analytical behaviorists never found a plausible way to draw the distinction between *action* (my raising my hand) and mere *movement* (my hand's rising) in observable, physical terms. Even had that problem been solved, there was a second one: the analytical project falsely presumed that every mental state had a proprietary behavioral profile, so that "wanting to attract attention" could be analyzed in terms of, say, raising one's hand. But a desire only produces action in connection with a belief (in this case, the belief that raising one's hand is a way to attract attention). The same desire combined with a different belief (say, that slapping a thigh is a way to attract attention) will result in one's slapping a thigh instead. Because the contents of beliefs and desires vary independently, reference to mental states is ineliminable.

Wittgenstein (1953) actually never endorsed any such project of analysis, but merely resisted what he thought was an excessively referential conception of the meaning of mental terms (as of language generally). He and Ryle argued that thinking of mental talk as referring to mental states that are the inner *causes* of behavior involved a misunderstanding of the "use" of mental talk, rather in the way it would be a misunderstanding of the use of "the average American" to take it as referring to an actual human being. In the 1940s and 50s, this general non-referential view of language gave rise to the movement in the 1940s and 50s known as "Ordinary Language Philosophy" (see "Ordinary Language Philosophy". In other hands it independently led both to "functionalist" conceptions of the meaning of many theoretical terms, especially mental ones, and to recent wariness about including reference in a linguistic semantics, as in Chomsky (2000), to both of which themes we'll return in §2.5.

This non-referentialist view has been regarded as particularly apt in understanding talk of sensations, as in "adverbialist" and "representationalist" theories of them (see Chisholm 1957, Dennett, 1991, Lycan, 1996, Rey, 1997a). However, the question of exactly *which* sorts of mental states can actually be thought of in this non-referential way depends in part on just what scientific explanations of human and animal behavior specifically require. At any rate, many philosophers have been sympathetic to at least the spirit of Quine's (1953a/61) "criterion of ontological commitment," according to which one should take as real all and only what is required by one's best explanation of a domain. And this takes us back, more than Wittgenstein and Ryle perhaps would have liked, to issues raised by the scientific work.

2.4 The Failure of Scientific Behaviorism

There is no question that Scientific Behaviorism vastly raised the standards of experimentation on the mind, introducing carefully designed, controlled experiments, the results of

which were subjected to painstaking statistical analysis. The experimental subjects were usually rats or pigeons, but because behaviorists were convinced that all learning, across species, was due simply to the action of the Law of Effect, they were sure their findings would generalize to the behavior of humans.

Ironically enough, however, the very rigor of Behaviorist experiments undermined their theory. For example, "latent learning" could be shown to occur, when the animal is sated, without reinforcement; as well as, "passively," without the animal emitting the supposedly reinforced response (rats could *run* a maze faster simply by wheeled passively through it in a cart, when they hadn't run it at all). And animals could improvise new appropriate behaviors not previously reinforced: rats trained to run a circuitous maze would take a novel short-cut when it was made available.[5] Indeed, one area in which experimental research has increasingly challenged, not only Scientific Behaviorism, but many traditional philosophical speculations about animals is with regard to animal cognition of the sort involved not only in navigation, but in tool use, planning, and even the understanding of other animal's mental states (see Gallistel, 1990, and especially Clayton et al., 2006, for surprising data about scrub-jay's apparent monitoring of each other's minds!).

2.5 Chomsky

Karl Lashley (1951) had raised a problem for Behaviorism about hierarchical structures in planning, a problem that was shown to be particularly profound by the revolutionary research program on the syntax of natural language initiated by Noam Chomsky (1959, 1965, 1968/72, 1980, 1981, 2000) and continuing to the present day (see the chapter entitled "Linguistic and Philosophical Methodology". Chomsky (1959) offered both a searing critique of the behaviorist account of language acquisition, and an unabashedly cognitivist alternative of his own. In his famous "Poverty of the Stimulus Argument," he argued that, where the behaviorist thesis was sufficiently well-defined (which he points out it often wasn't), it will be incapable of explaining how a normal child could acquire a competence to understand the virtual infinity of complex structures of a natural language on the basis merely of associative chains produced by conditioning to the physically characterized stimuli it receives in its first several years.

To provide just a few quick examples of the difficulty: consider, first, the following two ordinary sentences of English:

(1) John is eager to please.
(2) John is easy to please.

Utterances of these two sentences are as nearly *physically* alike as two different sentences can be. Nonetheless, native speakers of English construe them quite differently, understanding that "John" is the agent of the pleasing in (1), but the patient of the pleasing in (2).

[5] See Brewer (1974) and Rey (1997b: ch. 4) and for many further examples and discussion of some of the Behaviorist's responses.

Speakers' grasp of this fact is evident in their judgment that, while sentence (3) is acceptable, sentence (4) is not. (Note: a * indicates unacceptability.)

(3) John is eager to please Sally.
*(4) John is easy to please Sally.

Or consider a simple example of pronoun co-reference. In an utterance of (5), the speaker can normally intend "he" to refer either to John or to someone else:

(5) John hoped he would win.

However, this is not true in (6):

(6) He hoped John would win.

Here, a speaker must normally intend "he" to be anyone except John. (We say "normally," since a speaker could decide to deliberately flout a rule, as when people sometimes refer to themselves in a third-person way.)

This latter example is particularly interesting since the relevant rule[6] seems to be respected even by three-year-olds (see Crain and Thornton, 1994/2006), who were obviously never explicitly taught such a rule—few non-linguists are ever conscious of it—and which, again, involves some kind of appreciation of structural facts that are not audible in the phonological stream. Such examples—and linguistic textbooks can supply indefinitely many more—begin experimentally to vindicate the poverty of stimulus argument, which is further reinforced by the observation that, unlike the linguists, children have no access to anything like the full range of their own language, and, interestingly, almost never receive any of the "negative" data (telling them when certain sequences of words are ill-formed) on which linguists routinely rely: examination of corpora (see, e.g. Brown, 1973) indicate that virtually the only corrections children receive are for the *truth* of what they say, almost never corrections of grammatical errors of the sort indicated by our "*"s—which in fact neither they nor anyone (aside from linguists) ever produce. It is extremely hard to see how toddlers could possibly have acquired their competence from experience alone in the ways that the behaviorists claimed (although there continues to be work on sophisticated statistical techniques that their proponents think may help; see, e.g. Lappin and Shieber, 2007). Resurrecting what most psychologists and analytic philosophers at the time thought was a hopeless Rationalist doctrine, Chomsky forcefully argued that people's knowledge of the rules of grammar must be largely *innate*, in the form of a system of "Universal Grammar," or "UG", of which familiar spoken languages are simply specific variants, involving (in the terms of Chomsky, 1981) the setting of specific parameters that UG leaves open.

[6] Roughly, that a pronoun can't pick up its reference from a dependent clause that it dominates (technically, "c-commands").

Chomsky's program in linguistics became a main inspiration of the "cognitive science revolution" that in the late 1960s began sweeping psychology generally (see §4.3). Several methodological points emerge from it that are of considerable philosophical importance:

(i) *Idealization:* Perhaps one of the most important contributions Chomsky made early on was to vigorously challenge the behaviorist's assumption that the appropriate subject for psychology was the explanation of *behavior*. This was an assumption that was a consequence of the Positivist's insistence on the primacy of "observational" data in science, and its skepticism about "unobservable" theoretical posits. But Chomsky argued that this positivist view was hopeless as a conception of science generally, since the actual behavior of most large objects is the consequence of interaction of a great many separate "systems." No one expects there to be a science of the motion of leaves in wind and rain; rather, there will be sciences of, for example, aerodynamics and hydrodynamics, which together, and combined with facts about the shapes and composition of leaves, might explain why particular leaves move as they do. Similarly, animal behavior is the result of a massive interaction among, for example, its abilities, beliefs, desires, reasoning abilities, habits, proclivities, and available energy. The only explanatorily tractable approach to this enormous complexity is one that tries to discover what the different components might be, about which one might develop some systematic theoretical insight. In terms Chomksy introduced that have become enormously influential in psychology, theorists should be interested, *inter alia*, in a system's "competencies," not its actual "performance." And he thinks the evidence strongly suggests that grammar is just such a component competence. Speakers may enjoy a perfectly good competence with the potential infinity of English sentences, even though they may perform utterances of only a tiny fraction of them (and even a smaller one if they take vows of silence). Indeed, a theory of their grammatical competence may predict that they would be competent to understand a sentence consisting of a billion disjunctions of "Fish swim," even though, of course, no one will never utter such a thing. Contrary to the tenets of Positivists and Behaviorists, an "unobservable" underlying competence is an entirely apt object of linguistic study, posited to explain actual linguistic performance in interaction with indefinite numbers of other subsystems.

(ii) *Intuitions:* This interest in isolating the underlying competencies of a system also distinguishes a Chomskyan interest in linguistic intuitions from a standard philosophical interest in intuitions.[7] Whereas philosophers are often after merely a "systematization" of intuitive verdicts about some domain, Chomskyans are interested in *explaining* them, and discounting ones that are likely to be the result of interaction effects; and the explanation may well advert to states and processes not available to armchair intuitions.

[7] The status of intuitions both in linguistics and philosophy has been highly controversial; see Schütze (1996), Devitt (2006, 2013), Antony (2003), and Rey (2013b), as well as the chapter entitled "Intuitions".

For example, most English speakers would at least initially find unacceptable:

(7) The man that the dog that the lady loved bit died.

However, it is an empirically arguable question whether this is due to a genuine constraint of grammar, or to some problem with the organization of memory. After all, people have no problem with:

(8) This is the lady that loved the dog that bit the man that died.

But it's hard to see why a grammar that allowed (8) wouldn't allow (7). The problem with (7) is arguably due to some problem with processing the "center embeddings" of the clauses, due perhaps to a problem in the structure of short-term memory, a problem that doesn't arise in the right embedding of the nearly equivalent (8). But the structure of short-term memory is reasonably regarded as independent of the theory of grammar—and is not something that can be determined by either introspection or *a priori* reflection.

(iii) *Internalism*: Chomsky's theories are largely "internalistic," concerned only with the (often quite abstract) character of internal computations and the innate endowment that constrains it, largely independently of the environments a human might inhabit. He (2000) quite pointedly argues that linguistics should not be concerned with the relations of the mind to the *external world*, particularly of the sort discussed by philosophers when they offer theories of truth and reference. Here he is not only harkening back to traditional Cartesian and Kantian interests in the ways in which the mind structures whatever input it receives, but also, somewhat surprisingly, is taking up some of the non-referential suggestions of Wittgenstein that we have discussed. At any rate, he thinks that issues of truth and reference as they pertain to sentences of *natural language*, as opposed to a language of science, are too bound up with complex and highly variable pragmatic issues surrounding the *use* of natural language for them to be theoretically tractable, and so have no place in a serious *linguistic* semantics (one can get a feel for his skepticism here by reflecting on just what the real world "referents" might be of ordinary terms like "the sky," "rainbows," "Joe Six Pack," and "the inner track that Raytheon has on the latest missile contract"; see his 2000:135).

Chomsky, in short, revitalized a realistically mentalist, internalist and rationalist conception of psychology by presenting strong empirical arguments for the reality of elaborate cognitive structures, and for the necessity of innate knowledge for children to acquire linguistic competence. But at the same time, the conception is physicalist in its presumption that these structures are somehow realized in the human brain (which is why Chomsky often speaks of his theory as a "biological" account of language).

3. CRTT AND AN LOT

Simultaneous with the gradual demise of both forms of Behaviorism in the 1960s, there arose a scientifically attractive approach to the very inner mental processes that all behaviorists had tried to avoid, namely, a Computational/Representational Theory of Thought ("CRTT"). This was due largely to the development of computers as a result of Alan Turing's celebrated work applied to human problem-solving, as in Newell and Simon (1972). Fodor (1975:chap 2) argues that standard decision theoretic models of planning, and hypothesis—testing models of early vision and of concept acquisition, all presuppose a sufficiently expressive system of representation, a "language of thought" ("LOT"): a syntactically structured, semantically valuable, causally efficacious system of representation encoded in the brain (see also Pylyshyn, 1984).

An important feature of Turing's proposal was that the states of his machines essentially involved relations not only between input and output, but also *among themselves*. This led to "functionalist" theories, which characterized mental states in terms of stimuli, responses, and other mental states (Putnam 1960/75), a proposal that was developed in quite a number of ways (see Block, 1978/80, Shoemaker, 1981, and Rey, 1997b: chs 6–7). "Analytical functionalism" (Armstrong, 1968, Lewis, 1972) is really an extension of Behaviorism, purporting to capture the semantics of ordinary mental terms, and simply allowing the definition of a host of terms all at once, instead of each one separately. Lewis combined this technical idea with a proposal to extract the definitions from everyday "platitudes" about mentality and behavior.[8] In contrast, scientific or "psycho-"functionalism" is not a thesis about the ordinary meanings of mental terms, but about the actual nature of mental states, and proposed to draw from empirical psychological theories, the terms of which might differ markedly from those employed in everyday discourse (see Lycan, 1996). A psychofunctionalist theory, then, might posit any number of "unobservable" states as real causes of behavior, in keeping with the idea denied by behaviorists that what matters in psychology is not merely how a system behaves, but how the behavior is produced.

CRTT would seem also to be motivated by the aforementioned work of Chomsky. At least in many places (e.g. 1965: 30), Chomsky sketches his positive account of language acquisition in terms of a hypothesis confirmation model, whereby universal grammar makes available a limited set of possible specific grammars, and the task of the child is to select among them on the basis of the input provided by ambient speakers. He frequently refers to such a procedure as a "computational/representational" model, comparing it to the way a missile might be guided by internal computations over representations of its paths and targets (see 1980). However, it isn't clear that he intended these remarks literally,[9] even though they were certainly taken literally by many of his followers, particularly Fodor (1975: 58).

[8] This proposal was generalized by Frank Jackson (1998) and others into what has come to be called the "Canberra" plan for analyzing concepts generally.

[9] See articles of Egan, Gopnik, and Rey, with Chomsky's replies in Antony and Hornstein, 2003, as well as Collins, 2009.

A problem for computational models of the mind is that of accounting for the *meaning* or *content*—the "psychosemantics"—of the computational states, a problem that has often been under-appreciated by computer scientists. This is the hoary philosophical problem of "intentionality" or the problem of explaining the property by virtue of which a symbol or state of the brain is "about" some subject matter, for example a thought is about the number 7, or a walrus, or the Big Bang. The issue can be ignored by computer scientists largely because the intentionality of an *artifact*—for example, a computer the scientists have built—is provided for free: the programmer gets to say what the states *are* about, for example chess, or a war, or a conversational exchange. The intentionality of an artifactual machine is *derived* from the intentionality of its makers or users. But such stipulations have no place in using computational models to understand natural objects, like animals and human beings. In this case, one needs to supplement a computational model with a theory of the content of the representations over which the computations are defined.

For many philosophers and psychologists, the content of states of the mind or brain is determined by the inferential or "conceptual" roles those states play in reasoning: for example, conceptual content might be "derived from experience," or from sensorimotor primitives, in the style of traditional empiricists like Hume (see Prinz, 2002, and Barsalou, 2010); or it might be constructed from "prototypes," as in Rosch (1978/99) and Smith and Medin (1981); or be captured by "meaning postulates," as in Carnap (1947) and Hale and Wright (2000). But there have been a number of difficulties noted about such approaches, the most central one being the serious challenge raised by Quine (1953b/61) of distinguishing conceptual roles that are constitutive of meaning from simply tenaciously held beliefs (is "Bachelors are male" a matter of meaning, or simply a firmly entrenched commonplace? What makes the difference? Fodor (1998) presses essentially the same point against recent conceptual role theories of meaning in linguistics and psychology (see §4.2, Rey (2013a) and the chapter called "Analytic–Synthetic and A Priori–A Posteriori" for further discussion). Secondly, as Devitt (1996) emphasizes, any conceptual role theory would seem to have to "bottom out" somewhere in primitives not determined by role, but, arguably, by some sort of causal relation to the world.

In any case, since the 1970s, there has been increasing interest in various "externalist" causal theories of content. Stampe (1977), Dretske (1981, 1988), and Fodor (1987, 1991) pursue "informational" models, whereby a state or symbol means p iff it covaries with p in specific sorts of ways. Cummins (1996) and Gallistel and King (2009) add to such a theory a further constraint: that there be homomorphisms between a system of representation and the system it represents, as seems to arise in the case of the vector algebraic representations exploited by animals in navigating. Pursuing a more explicitly biological approach, Ruth Millikan (1984), Fred Dretske (1988), David Papineau (1987), and Karen Neander (2012 and forthcoming) have defended "teleosemantic" proposals, which seek to base a theory of content in biological functions. Tyler Burge (2010) also emphasizes the role of a creature's biological needs in its niche, although without an attempt to provide a fully fledged theory of content.

Methodologically, it is important to notice how these "naturalistic" approaches to content shift the focus of a theory of meaning from its traditional "armchair" status, where, as we mentioned, philosophers have relied merely on their intuitive verdicts about how they would describe possible cases. Although such verdicts are not irrelevant to a CRTT, the focus of attention in cognitive science, as in linguistics, has become the needs of a naturalistic *explanation,* not only of such intuitions, but of the rich cognitive processing of many animals. Hence in all the work on natural intentionality there has

been much discussion of ascription of content to systems less complex than a sophisticated philosopher in an armchair, for example, frogs darting their tongues at flies; ants navigating in the desert; bees dancing in a way that indicates sources of nectar; visual systems that seem to represent spaces, lines, surfaces, and simple material objects.

Some philosophers and cognitive scientists have objected to the "Cartesianism" of Chomsky's, Fodor's and other mainstream computationalist approaches to mentality, which treat mental processes largely in abstraction from the rest of the body and the surrounding environment. Advocates of "embodied cognition" (e.g. Robert Rupert, 2009) argue that mentality must be understood in terms that include the extra- cranial body; and advocates of "extended cognition" (e.g. Clark, 2011; Wheeler, 2011) say that mentality reaches out into the extra- cranial world, to incorporate notebooks and other mental prostheses (see Adams and Aikawa, 2010, for critical response). O'Regan and Noë (2001) have articulated and defended the "enactive" approach, according to which perceptual states involve physical action. Ned Block (2002a) has accused this last view of confounding causation and constitution: the fact that feedback from the body and physical activity in an environment is *causally* necessary for the proper development of perception does not imply that the capacity to perceive is *constituted* by that activity.

4. Cognitive Science

4.1 Cognitive Architecture

For those who accept the general idea that the human mind is a computational device, the question arises: what *kind* of computational device? There are two central questions, to which philosophers and cognitive scientists have devoted a great deal of attention:

(i) is the mind classical or connectionist in structure?
and
(ii) are mental operations modular or general in character?

4.1.1 *Classicism vs. Connectionism*

The first question concerns the kind of computation the mind performs. As we mentioned, Fodor (1975) argued that the mind must be presumed to be a kind of digital computer, with structure-sensitive operations defined over a symbolic system, a "language of thought," a conception of mind that became known as the "classical" computational model. However, this classical model was almost immediately challenged. Advocates of an alternative model of computation, "connectionism" (e.g., McClelland et al., 1986, and Smolensky, 1988), argued that the symbol-and-rules architecture of the classical model was not biologically realistic, and so likely not the way the human brain was organized.

A connectionist system, or "neural network," consists in a number of highly interconnected units, usually arranged in layers: an input layer, an output layer, and a variable number of intermediate layers, the last containing what are called "hidden" units. Whether a given unit fires or not depends upon two factors: the set of units to which it is connected, and the strength of the connections, which can be either positive or negative. These

connection strengths can be modified over time, as a function of the patterns of activity of the activating units, or of feedback from the systems receiving signals from the output layers. Proponents of a connectionist architecture argued that the neural network provided a much more plausible idealization of human neurology than did a classical architecture.

Intense controversy ensued, and continues to this day. The model we have just described seems simply to be a kind of behaviorist, associationist model, only "taken inside," and consequently much of the philosophical controversy mirrors the empirical controversy surrounding Scientific Behaviorism about the adequacy of associationism as a general theory of mind. Everyone agrees that a connectionist architecture does well at modeling the learning and deployment of statistical regularities; the question is how much of human thought can be explained in those terms? Just as Chomsky challenged Skinner's account of language acquisition by pointing to structure-dependent rules, Fodor and Pylyshyn (1988) pointed to what they argued was the inability of connectionist architecture to model structure-sensitive learning and reasoning. They claimed that human thought was essentially *compositional* (understanding a complex thought is a function of understanding its parts), and *systematic* (e.g., one think John loves Alice iff one can think Alice loves John), and that purely connectionist systems could not model these features. Connectionists responded to these criticisms in various ways: some argued that Fodor and Pylyshyn were wrong about these features, while others accepted that thought had these features, but argued that connectionism could model them—without being merely an implementation of a classical proposal. (Rey 1997b: §8.8, §9.4) cites some dozen phenomena that are challenges for Connectionism, but are easily explained by a classical view.)

The boundary between empirical psychology and philosophy has been especially porous in these discussions. McLaughlin and Warfield (1994) discussed the computational properties of classical and connectionist systems, showing that many of the claims made by Connectionists about the speed of connectionist networks and the "brittleness" of classical systems were not confirmed by the actual track records of existing programs. Marcus (2001) argues that such extra-empirical "virtues" of hypotheses such as simplicity cannot be used at this early stage of understanding to decide between connectionist and classical models.

4.1.2 *Modularity*

The second major question about cognitive architecture concerns *modularity*. In 1983, Fodor published an enormously influential book, *The Modularity of Mind*, in which he argued that there were two types of mental systems: "modular" ones, which included perceptual systems and a system for understanding speech, and "central" ones, responsible for belief fixation, planning, and conscious inference. Per the traditional "hypothesis testing" views of early visual processing, Fodor regards modular processes as computational, but ones that are typically fast, automatic, bottom-up, domain-specific, informationally encapsulated, and neurologically localized.[10] Fodor's view that perception is informationally encapsulated is opposed to the view that perception is "cognitively penetrable," or

[10] Think of the ways in which visual illusions, like the Müller–Lyer, exhibit these properties, for example, being immediate, and persisting even when the person knows better; and how difficult it is to hear one's native language as mere noise.

sensitive to the subject's background beliefs or interests. Jerome Bruner (1957) championed this view that came to be known as "New Look" psychology, a view that was extremely influential, both in that field and especially in philosophy of science (see Hanson, 1958/ 2010, Kuhn, 1962/ 2012) from the 1940's up until the publication of Fodor's book. Although it is no longer the consensus view, New Look psychology is enjoying a resurgence among experimental psychologists. For example, Bhalla and Proffitt (1999) have presented data that they say show that subjects see a distant hill as being steeper when they are wearing a heavy backpack than they do when unencumbered. In reply, Firestone and Scholl (2014) have argued that this work is based on a fundamental confusion, what they call the "El Greco Fallacy" after a notorious art history hypothesis that El Greco's distorted figures were the product of an astigmatism. The fallacy is that, if El Greco did have an astigmatism, he would "distort" the canvas in the same way, so that the mapping from figure to canvas would be preserved. Firestone and Scholl argue that advocates of top- down effects in perception make the same mistake in trying to demonstrate a subject's visual distortion by having the subject gauge it on a reference object (which should be similarly distorted).

The idea that perception is cognitively penetrable has recently been resuscitated by some philosophers. Susanne Siegel (2010) has defended the view that learning affects the content of perceptual experience by making available to perception high- level or theoretical properties. She claims that such top- down processing is the best explanation for the phenomenological difference many subjects report between, for example, looking at an unfamiliar script (like Cyrillic) before and after learning a language that utilizes that alphabet. Another view that, arguably, treats perception as cognitively penetrable is the view that we can perceive "affordances," or ways in which objects can be used or manipulated by us. The view was defended by psychologist James Gibson, and some philosophers defend similar views today, for example, Ruth Millikan (1995). In contrast to modules, Fodor holds that, "central systems," are slower, non-automatic, holistic, domain-general, and informationally promiscuous: they have a vocabulary general or flexible enough to express an unbounded number of distinct concepts, and these computations may draw from any information the organism possesses, a feature Fodor calls "isotropy". These facts would seem crucially to underlie the ability to apprehend relevance across subject boundaries, as in everyday problem-solving (e.g. using a tie twist to repair one's glasses), as well as in science (think of Newton, unifying terrestrial and astronomical dynamics under a single theory; see Antony, 2003, for other examples).

Fodor (2000) argues that it is difficult to see how this global isotropy could be modeled computationally, at least within the constraints of time and space that limit human cognition, and despairs of there being any computational account of central processes. He points out that the problem is related to the infamous "frame problem," or the problem of rendering explicit all of the *relevant* background knowledge an agent is able to deploy in acting or problem-solving. Ironically, for Fodor, CRTT might turn out to apply only to encapsulated modules, not to central reasoning (although the latter, for him, will still require the expressive power of an LOT).

Daniel Kahneman (2012) might appear to be proposing a similar "modular" view, when he proposes his "Two-Systems" theory of judgment. Although he officially denies that his view has ontological import—he says the terms "System 1" and "System 2" are merely abbreviatory for what may be several different kinds of system—he often encourages the interpretation that the two systems are real architectural components of the mind, and that appeal to one or the other "system" has explanatory (rather than merely descriptive) value.

At any rate, many philosophers have construed the theory in this way. Tamar Gendler (2008), for example, has argued that there is a distinctive type of System 1 cognitive state—a state she calls "alief"—that's analogous to System 2's belief, but which is implicated in such apparently irrational conditions as phobias and unconscious bias.

Gerd Gigerenzer's (2001) "bounded rationality" is yet another modular view. Gigerenzer identifies a host of cognitive shortcuts ("biases and heuristics") that give better results than standard reasoning models when applied in their domains of "ecological validity." For example, the "recognition heuristic" says to prefer familiar alternatives to unfamiliar ones. It works when the subject knows a little, but not a lot about the domain, and where fame (or notoriety) tracks the target property. So, for example, German college students were asked which of two cities had the higher population, Detroit or Milwaukee. Since few of them had even heard of Milwaukee , they chose, correctly, Detroit. Americans, who had heard of both cities, were at chance.

There have been several attempts by philosophers to deal with Fodor's problem of the apparent isotropy of human thought. One approach is to embrace his pessimistic conclusion that central cognition cannot be modeled classically, and to plump therefore for alternative architectures. Terry Horgan and John Tiensen (1994), for example, have argued that isotropy is a good reason to favor a connectionist model of central cognition, since the massive connectivity of a connectionist network can easily model the universal availability of information. Other philosophers, for example van Gelder and Niklasson (1994), have argued that Fodor has exaggerated the degree or the extent of isotropy in human cognition. Perhaps in practice, isotropy really only holds over relatively small chunks or families of information, assembled on an *ad hoc* basis to deal with specific questions or problems. If the hypothesis space in any particular task is limited, then it can be modeled by classical architecture. Arguably, neither of these strategies will work: connectionism implausibly makes *all* information always relevant, and the limited isotropy strategy fails to explain how *ad hoc* clusters of information are formed.

A third strategy, which has received a great deal of philosophical attention, is to posit more modules, carving out new domains of specialized thought and knowledge, and leaving less inferential work for central cognition to do. As we mentioned, Fodor argued for there being at least six modules: one for each of the usual five senses, plus another for language. But cognitive psychologists studying a range of distinctively human knowledge and behavior, and utilizing new experimental paradigms to test younger and younger children, have found experimental evidence of other domain-specific and relatively encapsulated bodies of information. These are either present innately, or else emerge extremely early and according to regular developmental patterns. Such modules have been proposed for face recognition, basic knowledge of objects (Spelke, 1998) and appreciation of other minds (Leslie, 1987, Frith, 1989, Baron-Cohen, 1996, Baillargeon et al., 2010) as well as for certain basic parts of morality (see §4.4).

There are, obviously, important foundational questions here, the leading one being: what does it mean to say that neonates have *knowledge* of physical objects, or human psychology? Developmental psychologists, psycholinguists, and vision scientists all help themselves to the notions of "knowledge" and "representation," and by and large leave it to philosophers to worry about what these notions mean. Chomsky's position on this has been adamant: there is no question about what it means to say that anyone "knows" a grammar beyond the question of the explanatory adequacy of the theories that utilize that notion. Devitt (2006), however, argues that the existing evidence only supports claims about skills and abilities, which, he argues, need not be based on any explicitly represented information (see Collins, 2008, for a reply). There has been a similar debate as to whether the phenomena that are

cited as supporting a "theory of mind" module are best explained in terms of an internalized theory, from which subjects draw inferences about the psychological states of others (so-called "theory" theory; see Gopnik, 1993), or in terms of some non-representational process, like simulation (e.g. Gordon, 2001).

It's important to note that most of the psychologists working on modular knowledge still accept that there is a role for central cognition to play. One particularly interesting suggestion is that central cognition plays a largely *integrative* role, providing a kind of *lingua franca* into which the output of modules with their own computational languages can be translated, and bringing to bear conceptual resources not available *within* a module. Susan Carey (2009) has developed such a theory to account for the development in humans of *numeracy*. Human infants, like many other mammals, have a couple of ways of representing quantity: there is an analogue system that can judge relative sizes of things, and there is a system that can generate up to six or seven individuative representations (proxy elements for objects in the world). What human beings can do that, apparently, no other animal can, is to somehow "bootstrap" these abilities into an abstract understanding of the numbers as an infinite series of elements generated and ordered by the successor function. Carey's hypothesis is that it is our innate language capacity that contains the conceptual resources to do this.

A different approach to modularity dispenses with central cognition altogether. The so-called "massive modularity" approach holds that cognition consists entirely in interacting modules. For Peter Carruthers (2006), for example, the appearance of executive function with respect to such processes as belief fixation and abductive inference is explained by one or both of two things: first, the existence of a common "global workspace," in which information from different modules can be exchanged, and second, by the operation of dynamic, non-intentional processes that direct energy throughout the cognitive system. Relevance relations, for example, are reflected in paths of least resistance between representations in various domains. It should be noted that for Carruthers and other massive-modularity theorists, a "module" need not display all the characteristics Fodor takes to be part of the module-syndrome. It's enough that they be domain- or function-specific, that is, that they operate only on certain kinds of input, or utilize specialized algorithms.

4.2 Cognitive Science and Epistemology

In two important papers, "Two Dogmas of Empiricism" and "Epistemology Naturalized," Quine (1953b/61, 1969) called for a methodological sea change in the philosophical study of knowledge. He argued that philosophers interested in the structure of knowledge should abandon attempts to give an *a priori* theory of empirical justification as they traditionally have done, and look instead at the actual conditions and processes that make human knowledge possible. Instead of asking how justification *ought* to proceed, we should ask how it *does* proceed. We should "just settle for psychology," an approach he called "naturalized epistemology." (Ironically, Quine didn't really follow his own advice: cleaving to Scientific Behaviorism to the end of his life, he largely disregarded all the work in cognitive science that we have outlined here; see Antony, 2000.)

The idea that the philosophical study of knowledge could and should be continuous with—or even discarded in favor of—the empirical study of perception and cognition scandalized many philosophers. Ernest Sosa (1983) and Jaegwon Kim (1994) protested that epistemology was fundamentally a normative project, and so could not be replaced

or constrained by a descriptive science. However, philosophers like Alvin Goldman (1986) and Hilary Kornblith (1994) saw empirical psychology, not as a replacement for epistemology, but as a welcome source of constraint.

Goldman's (1986) naturalism led him to propose a causal theory of knowledge, according to which the externalist notion of a reliable causal process largely replaced the internalist requirement of justification in the more traditional conception of knowledge. Other naturalistic epistemologists appealed to empirical work by Kahneman and others to argue that epistemology needed more realistic conceptions of rationality, deliberation, and self-knowledge than the tradition had assumed, especially if it meant dispensing useful epistemic advice to actual human beings.

Although it is beyond the scope of this chapter to review all of the work in epistemology that naturalistic philosophers have produced (see the chapter called "Philosophical Naturalism"), we'll simply mention a few issues where, ironically enough, a naturalistic approach has yielded challenges to elements of Quine's own critique of traditional epistemology.

(i) *Foundationalism*—The idea of sub-personal computational processes brings with it the possibility of unconscious but justificatory inferences from sensory data to perceptual belief. Combined with the thesis that sensory input systems are informationally encapsulated, we get a kind of vindication of epistemic foundationalism.

(ii) *Nativism*—Poverty-of-stimulus arguments, together with experimental demonstration of surprising cognitive abilities in very young babies constitute an empirical case for the existence of some forms of innate knowledge (and maybe a solution to the epistemological problem of other minds). (See Goldman, 2006, and Antony, 2004. But see Cowie, 1999, and Devitt, 2006, for empiricist challenges.) Although not quite a vindication of "first philosophy" as a method, such considerations do explain the ubiquity and utility of many of our *a priori* intuitions about the world.

(iii) *Analyticity*—Empirical work on semantics and pragmatics in natural language provides some evidence in favor of a principled distinction between "knowledge of meaning" and "knowledge of fact." For example, along the lines of Chomsky's non-referentialist suggestions mentioned earlier, Pietroski (2010) proposes that linguistics semantics is not concerned with truth and reference, but with "instructions" from the language faculty to the conceptual system about default constraints on the use of expressions. Combining this proposal with proposals of Fodor (1987) and Horwich (1998), Rey (2009) suggests that, if such default instructions provide the basic explanation of the use of an expression, on which other users asymmetrically depend, then they could provide a basis for analyticities (which, however, would not be immune to empirical revision, in the way that Quine worried). But see also Fodor and LePore (2010) for a surprising defense of Quine's original challenge, which they think has still not been met.

4.3 Evolutionary Psychology

One important motivation for massive modularity has been consideration of the possible evolution of human cognition. Stephen Pinker (2003) and Leda Cosmides and John Tooby

("C&T") (1992) have argued that problems of survival facing our hominid ancestors put pressure on cognitive evolution, leading to a variety of specialized cognitive structures in the modern human brain. For example, C&T have posited a "cheater detection" module, a mechanism that facilitates inference whenever rules of social exchange are involved. They point to the fact that subjects' performance on a certain logical inference task[11] is very poor when the problem is presented abstractly, but is significantly better when the problem is presented in the context of social exchange. Thus, for a generalization like "All cards with triangles on one side have red patches on the other," subjects will fail to see that a card with a green patch on its face is potentially a counterexample. If, however, a generalization of the same form is presented in the context of a story about social rules ("In society X, it is a rule that all cassava eaters must be married"), subjects recognize immediately that they must check the eating habits of unmarried men to see if the rule is being followed.

C&T argued that this pattern of results tells against the view that human minds are general-purpose logic machines, and in favor of the view that we have specialized mental mechanisms that facilitate reasoning in specific domains. A "cheater detection" module makes sense from an evolutionary point of view: we are a social species, and social species need rules to organize interpersonal behavior and to reduce conflict. But rules are useless unless they are enforced. Hence, there would have been, in the ancestral environment, selective pressure in favor of a mental mechanism that facilitated the recognition of the conditions that constitute cheating in a given case, pressure, in other words, towards the evolution of a "cheater-detection module."

C&T see themselves as arguing, not only for a particular conception of cognitive architecture, but for a new approach to the study of the human mind. Evolutionary psychology, they tell us, is "the application of adaptationist logic to the study of the architecture of the human mind." What C&T have in mind, it seems, is that consideration of the evolutionary challenges facing our ancestors could serve as a discovery procedure for new hypotheses about mental structure. In practice, this has not happened, largely because so little is actually known about either the psychologies of early hominids or the circumstances in which they evolved.[12] C&T contend, however, that we can learn about the evolutionary function of cognitive structures without knowing human prehistory by apprehending the "complex functional design" evident in the current human phenotype. They point to vision science as an example of the fruitfulness of this sort of approach.[13]

4.4 Moral Theory and Psychology

The philosophical study of morality has always involved, at least tacitly, empirical assumptions about human psychology. Ancient philosophers took it for granted that notions like virtue and duty were bound up with observable facts about human capacities and interests, and that moral and political theories had to be constrained by the conditions of human flourishing. In the modern period, views about human psychology and rationality visibly

[11] The "Wason Selection task": see Wason (1966).

[12] But see Anne Fausto-Sterling, 1997, for suggestions about data that do bear on evolutionary history, but that evolutionary psychologists ignore.

[13] One might note, however, that the kind of "function" at work in the methodology of vision science is synchronic, not historical.

shaped moral and political theory. Rationalists, like Spinoza, emphasized the normative importance of human freedom. This view found its fullest expression in the work of Kant, who took autonomy—our ability to set rules and ends for ourselves—to be the foundation of both human moral value and of rational agents' duties to each other. Empiricists tended to place human affective capacities at the center of ethical theory: Hume's sentimentalism focused on sympathy, while Bentham's and Mill's utilitarianism made human happiness the foundation of moral obligation. Social contract theorists through the ages (Plato, Hobbes, Locke, Hume, Kant, and Rawls) have sought to ground justice and other political values in rational self-interest.

One role, then, for empirical psychology to play in the study of normativity is to fill out, and then to test experimentally the somewhat inchoate pictures of human capacities and propensities on which these and other philosophical theories rely.

Consider, for example, the normative theory known as "virtue ethics." According to this view, a person's behavior is morally good to the extent that it springs from *virtue*. Virtues, like honesty, courage, and compassion, are character traits distinctive of paradigmatically virtuous persons. Such traits are presumed to be relatively stable features of personality, predictive of behavior across a wide variety of circumstances. Gilbert Harman (1999) and John Doris (2002), however, argue that the existence of character traits in this sense has been called into question by the social psychological research of the last half-century. beginning with the infamous Stanley Milgram experiments in 1963. Milgram (1974) found that ordinary, presumably decent people could be shockingly obedient to authority, showing willingness to inflict high degrees of pain on innocent others (as the subjects believed), just because they had been asked to do so by the experimenter. Other studies found that people's willingness to aid strangers in (apparent) distress could be predicted by such ethically irrelevant factors as whether they had just found a coin (Isen and Levin, 1972), and whether there was an unpleasant noise in the background (Mathews and Cannon, 1975). Harman and Doris argue that such findings show that ethical theory should be *situation-based* instead of character-based: that it should focus on discovering the kinds of situational factors that promote or inhibit morally good behavior. They concede that there will be individual differences in behavioral dispositions, but contend that these "fine-grained" character traits will be too complex and idiosyncratic to support useful generalizations.

Defenders of virtue ethics have replied that their situationist critics are attacking a strawman, and that the conception of "character trait" operative in virtue theory is far more subtle than the critics presume. Rachana Kamtekar (2004), for example, says that virtue is a complex set of dispositions, involving cognitive and affective states as well as actions. A sympathetic person may not always act rightly, but she will feel remorse or shame if she doesn't. Julia Annas (2005) has argued that virtue is manifest primarily as a "firmness in intelligent deliberation", and only indirectly in behavior. It is worth noting in this connection that all of Milgram's subjects, including some of those who were willing to inflict the maximum degree of "pain," evinced a great deal of distress, trembling, crying, and biting their lips and hands. Some asked to be allowed to switch places with the other "subject." And a sizeable minority of the subjects (35%) did, at some point, refuse to obey the experimenter (Perry, 2012).

Apart from virtue ethics, the two dominant strains of thought in normative ethics are consequentialism, which says that the moral value of an action depends on the goodness of its consequences, and deontology, which holds that morally right action is action performed in accordance with moral rules. There is considerable literature defending both views, but some recent empirical work bears on it.

Joshua Greene (2013), for example, believes that consequentialism is correct, and that the moral intuitions that are appealed to in support of deontological theories can be explained away as psychological artifacts. Greene focuses on the results of a pair of thought-experiments devised by Philippa Foot (1967). In both cases, there is stipulated to be a runaway train which, unless something is done, will run over and kill five people tied to the track. In "Trolley," subjects are asked if it would be morally permissible to divert the train to a different track, where only one person is tied to the track. In "Footbridge," subjects are asked if it would be morally permissible to push a person large enough to stop the train off ~~of~~ a footbridge into its path, killing the large person, but saving the five. Typically, subjects judge that it is all right to switch the train, but not all right to push the large person.

This pattern is puzzling from a consequentialist point of view, since the consequences in both cases seem to be the same: five saved, one lost. Deontologists, however, explain the observed pattern of judgments in terms of subjects' adherence to a moral rule that makes a distinction between merely forseeing harm and intending it, the so-called "Rule of Double Effect." Trolley obeys this rule, but Footbridge violates it, since in Trolley the death of the one is foreseen, but not intended, whereas in Footbridge, the death of the one is a necessary element of the plan to save the five.

Greene, however, has a different analysis of these judgment patterns. He believes that human moral commitments are fundamentally consequentialist. If so, we would realize, at some level, that the two conditions are morally equivalent and that we should give the same answer in both cases. The second case, however, involves "up-close and personal" aggression against the large person. Because we have an evolved aversion to laying hands on another person, consideration of the Footbridge case causes us a great deal of emotional distress, which motivates us to seek a rationalization for not pushing the large person. The rule we come up with—"Don't use another person as a mere means"—is, then, simply a *post hoc* rationalization. Greene supports his view with brain-imaging studies of subjects engaged in deliberation about these cases. The studies show significant amygdala involvement, indicative of emotional engagement, in subjects contemplating Footbridge, but not Trolley.

Greene's conclusions have been challenged not only by philosophers (Berker, 2009), but also by psychologists. Saunders (2014) has recently presented a version of the Trolley/Footbridge cases to children aged 3–4, involving a thieving squirrel intent on eating some children's cookies. In the "Trolley" condition, the subject can prevent the theft of five children's cookies by erecting a barrier, with the consequence that the squirrel will steal one cookie from a child on the other side. In the "Footbridge" condition, the subject can prevent the theft of the five cookies only by taking the sixth child's cookie and feeding it to the squirrel. The young subjects' judgments pattern was just like the adults, despite there being no "up-close-and-personal" aspect to the rejected option.

Deontological theories give strong emphasis to the *reasoning* that goes into a moral judgment. According to such theories, it is not enough for an agent to conform to the pertinent moral rule: the agent must act out of an appreciation of that rule. This requirement, however, seems at odds with many of the facts about moral judgment. Many, if not most, of the moral judgments we make come quickly, without any discernible deliberation. Moreover, many of us are subject to a phenomenon called "moral dumbfounding," in which we report strong moral intuitions about a case, but then can supply no justification whatsoever for our judgments. For example, many people are adamant that brother–sister incest is morally wrong, even in cases where the usual grounds for objection have all been removed (there is no chance of pregnancy, neither sibling has been coerced or deceived, etc.). Jonathan Haidt (2001) believes that such examples show that moral judgment is not fundamentally cognitive, much less rational, but is rather more like perception. Like Greene, he believes that the "reasons" people sometimes give to justify their moral judgments, are post hoc rationalizations, and do not reflect the actual causal determinants of the judgment.

Defenders of deonotology, as well as consequentialists who think that rational defensibility is essential to sound moral judgment, have resources to meet this challenge. One is to appeal to Kahneman's "two-systems" approach to everyday cognition that we discussed earlier. Snap moral judgments can be the result of associative connections or "good-enough" heuristics that happen to be applied in cases that demand more careful consideration. In this respect, the analogy with perception is not inapt: just as modularists have emphasized how perception differs from belief fixation (recall the Müller–Lyer illusion, section 4.1.2), moral associations or heuristics may deliver judgments that a responsible moral agent will, on reflection, override).

Indeed, a number of theorists have proposed conceiving of moral thought on the analogy of Chomsky's distinction between competence and performance that we discussed earlier. This analogy was first suggested by John Rawls' in his (1971) A *Theory of Justice*, and has recently been pursued by John Mikhail (2011) and Susan Dwyer (2009). They posit a "universal moral grammar," analogous to Chomsky's UG, that sets the parameters for specific systems of moral permissions and prohibitions that are triggered in neurotypical human beings upon casual exposure to mature moral reasoners in a given culture. And there is some evidence, both from studies of small children and from research on nonhuman primates, that some aspects of moral competence are innate in human children (see Brosnan and de Waal's, 2003, and Baillargeon et al., 2010). But just as people's actual speech may not always reflect their grammatical competence, their moral judgments and actions may not always reflect their underlying moral competence. Moral judgments may just be the noisy result of interactions of it with memory, perception, and—especially in the case of morals—emotion, beliefs, and motivations.

Empirical psychology, both cognitive and social, has had an enormous impact on practical ethics and political philosophy. To cite just one example: philosophers working on race and gender now have at their disposal a rich body of work on the nature, etiology, and effects of racism and sexism (e.g. Steele, 1997, on stereotype threat; Banaji et al., 2013, on implicit bias; S. Leslie, 2015, on the cognition of prejudice), as well as promising research on strategies for combatting pernicious social attitudes (Walton, 2014).

Feminist philosophers have questioned whether gender plays a role in moral reasoning. Carol Gilligan (1982) famously identified two different "voices" in the moral reasoning of men and women: an impersonal, abstract "justice" perspective, and an empathetic, situated "caring" perspective. Although Gilligan's original claim that there was a gender difference in children's deployment of these perspectives has not held up (see Walker, 1984), her work has still inspired new approaches to moral problem-solving that foreground the interpersonal skills and concerns associated with women's caring work (see Ruddick, 2002, as well as the chapter called "Feminism").

Moral philosophers have drawn on empirical studies in psychology in many more ways than we have space here to survey. Interested readers are advised to consult Sinnott-Armstrong's, 2008, three-volume reference work on moral psychology for a more comprehensive treatment of the topic.

5. Two Specific Problems where Psychology Affects Philosophy

To make particularly vivid the increasing relevance of experimental psychology on philosophical issues, we want to mention two important problems that have been the focus of a good deal of recent discussion in both fields:

5.1 Consciousness and its Supposed Unity

Thomas Nagel (1974/79) and Frank Jackson (1986/92) raised a challenge for the kind of scientific psychology we've discussed here, calling attention in different ways to what Joseph Levine (1983) called the "explanatory gap" between the kind of objective knowledge we might have of a brain and the special "subjective" knowledge we might have, or want to have, about the associated states of a mind. It appears that, even if there were a complete objective description of an entire brain, it wouldn't (as Levine nicely puts it) "upwardly necessitate" the descriptions that might be provided subjectively, for example for Nagel, of "what it would be like to be a bat," or, for Jackson, experience of red. This fact seems to stand in sharp contrast to the way that standard laws and micro-descriptions in physics and chemistry plausibly promise to upwardly necessitate virtually any other non-mental phenomenon, say, why water expands when it freezes. Given the laws, geometry and micro-descriptions, it follows that frozen water *must* expand, whereas it can seem always to be an open question, say, whether a certain brain state corresponds to an experience of red, as opposed to green; or even whether there's a conscious state at all (the brain might be that of a "zombie"). Putting to one side the avalanche of purely philosophical responses to the problem, there have been a number of empirical results that, while they certainly don't completely solve the problem, seem to bear importantly on its solution.

There are, for example, a number of surprising results regarding the timing of experience and the associated brain states, usefully assembled by Dennett (1991). In a much

discussed experiment, Kolers and von Grünau (1976) had subjects report their experience of a phi-phenomenon,[14] specifically, of two rapidly successive flashes of two dots, a red one on the left followed by a green one on the right, which subjects see as a single dot moving: the question was *when* did the dot appear to change color? Subjects reported that it did so *before* "it" arrived at the right. Apparently, the subjects *retrospectively* "filled in" the color of the apparently moving *spot* after their eyes saw the color of the right dot, even though it appeared to change beforehand. So just *when* did the appearance of the spot midway as green occur, before or after its appearance as green on the right? An experiment by Geldard and Sherrick raises a similar problem: subjects retrospectively perceived taps on their arm, even the initial ones, as due to the movement of a small animal. There are also temporal puzzles surrounding the "conscious" initiation of a movement that *seems* to be preceded by the neural events directly responsible for the movement, as in Libet (2004). Dennett (1991:chps 5–6) argues that these results tell against what he takes to be an assumption, shared by dualists and materialist alike, of a "Cartesian Theatre," or the assumption that there is an internal mental stage in which all the important mental events "come together and consciousness happens" (p. 39).

That consciousness might not be quite as unified a phenomenon as we ordinarily suppose is even more dramatically illustrated by the startling "split-brain" results. In the 1940s, a procedure was developed for the relief of some particularly severe cases of epilepsy that consisted in the severing of the *corpus callosum*, or the fibers connecting the right and left hemispheres of the brain. The procedure did relieve the epilepsy, but turned out to have a surprising consequence. Michael Gazzaniga et al. (1962) carefully segregated the stimuli causing signals to the different hemispheres, so that the left hemisphere (typically the seat of most linguistic abilities) didn't have access to the signals sent to the right. The result was that, at least for the duration of the experiment, the original person seemed to be divided into two: most (although not all) personal cognitive capacities—thinking, comparing and retrieving objects, were sharply divided between the two hemispheres in such a way that, while each hemisphere seemed to embody fairly whole and adept people, neither was able to integrate its functioning with that of the other. The right hand, so to say, didn't know what the left hand was doing.

These experiments raise bewildering problems about personal identity and the unity of consciousness: are these people with severed *collosa* to be counted still as one person? Or as two, corresponding to the two hemispheres? Or as *three*, one composed of the other two? And if so, how many people were there before the operation, or in normal cases? And what about further divisions that might be surgically drawn, say, between various reasoning faculties, as in the Freudian unconscious, or in Kahneman's aforementioned two "systems" of reasoning? How are we to count persons, and understand the supposed "unity of the person" in the light of these empirical results? See Nagel (1971/79), Parfit (1984), Marks (1981), Gazzaniga (1995), Bayne (2010), and Schechter (forthcoming) for discussion.

[14] This is the phenomenon whereby subjects see apparent motion if two similar images are flashed in rapid succession, as in movies.

These latter questions dramatize even further Levine's explanatory gap, raising what Chalmers (1996) has called "the hard problem" in the philosophy of mind, the metaphysical explanation of consciousness, and what Block (2002b) has called "the harder one": how one could ever *know* whether another creature or machine that satisfied the metaphysical explanation was actually conscious. But Block, himself, has said useful things: in his (1995), he reviews a great deal of empirical work that he claims conflates what he calls "A(ccess)-" with "P(henomenal)" consciousness, and claims it's the latter, not the former that presents the problem of the gap; and in his (2011), he argues that experiments of Sperling (1960) show that phenomenal consciousness involves sensory material beyond cognitive access. His discussion intersects in interesting ways with proposals of Jackendoff (1987) and Prinz (2012), according to which the contents of consciousness are shallow perceptual ones, on the model of the "2-1/2D" sketch proposed by David Marr (1982) as the output of early visual processing.[15] Prinz also assigns a crucial role to attention in consciousness, a view challenged by Kentridge, Heywood and Weiskrantz (2004). Something akin to the sensory proposal is shared by Peter Carruthers (2011), when he considers the question of what *access* we have to our own conscious states, which brings us to our second problem:

5.2 "Blind-Sight" and "Self-Blindness"

In a well-known article, the philosopher Sydney Shoemaker (1996) presented arguments against the possibility of "self-blindness," or the inability of someone, otherwise intelligent and possessed of mental concepts, to introspect any of her concurrent attitude states. However, along the lines of the "theory theory" we mentioned earlier, the psychologist, Alison Gopnik (1993), and, more recently, the philosopher, Peter Carruthers (2011) argue that there is substantial evidence that so-called introspection is largely *interpretive:* the supposed introspection of at least one's propositional attitudes (such as belief or desire) is not "direct," as it is in the case of sensory perception, but an "inference" in the way it is in the case of our beliefs about the attitudes of other people; moreover, an inference that has the advantage in our *own* case of only our own sensory data and memories, and of the context we are in: that is, flying directly in the face of Shoemaker's claim, they are arguing we are all substantially self-blind! (Still more controversially, Rey, 1997b, suggests that even one's own self-attributions of consciousness could involve in a similar way an inferential imposition upon oneself of a certain, perhaps innate theory.)

Carruthers cites as evidence for his claim, experiments showing that people are often confabulating when they think they are introspecting. For example, they are unreliable about distinguishing genuine memories from rational reconstructions of the past; they can confuse vivid mental images for actual visual experiences; and they seem to be poorer than one would expect at monitoring the mental processes that can be independently

[15] The "2-1/2D sketch" is essentially a detailed left/right, up/down image, with only approximate indications of relative depth.

shown to responsible for their behavior: Brasil-Neto et al. (1992), for example, showed that in certain cases subjects can think that their decision to move a certain finger is the cause of their finger moving, when in fact the cause is an independent electrical signal that is not plausibly causing any such decision (see also the aforementioned Libet, 2004, as well as Wegner, 2002, for further examples; and for discussion of further introspective errors, Nisbett and Wilson, 1977, Ericsson and Simon, 1993, and Wilson, 2002). Particularly striking are also some of the split-brain cases, in which patients verbally report what seem to be introspected reasons for actions that are patently the consequence of stimuli available only to the non-linguistic hemisphere (Gazzaniga, 2000, discussed at Carruthers, 2011: 39–40).

There's actually a famous case that Carruthers doesn't discuss (he's concerned only with supposed introspection of propositional attitudes, such as belief and desire), and this is the surprising phenomenon of "blind sight," which involves a dissociation between introspective claims and what appears to be at least some serious kind of visual experience: subjects with damage to certain areas of their visual cortex can be shown to be sensitive to certain visual stimuli, such as the presence, shape and color of objects, despite their introspective assurances that they could see nothing! See, for example, Weiskrantz (1997), Rosenthal (1993), and Milner and Goodale (1995) for discussion.

A number of philosophers have defended the possibility of at least *some* special introspective knowledge of *some* of one's attitudes in the face of these results (e.g. Goldman, 2006, Nichols and Stich, 2003, Rey, 2013). Whatever the ultimate verdict about such cases, the issue seems clearly an intricate empirical one, not to be settled by armchair reflection or introspection alone. It may simply not be introspectible, whether we have special introspective knowledge or not.

Bibliography

Adams, F. and Aizawa, K. (2010), "Defending the Bounds of Cognition," in R. Menary (ed.), *The Extended Mind*, Cambridge: MIT Press, pp. 67–80.

Annas, J. (2005), "Comments on John Doris's Lack of Character," *Philosophy and Phenomenological Research*, 71: 636–42.

Antony, Louise M. (2000), "Naturalizing Radical Translation," in A. Orenstein and P. Kotatko (eds.), *Knowledge, Language, and Logic*, Boston Studies in the Philosophy of Science, Dordrecht: Kluwer Academic Publishers, pp. 141–50.

Antony, Louise M. (2003), "Rabbit-Pots and Supernovas: On the Relevance of Psychological Data to Linguistic Theory," in Alex Barber (ed.), *Epistemology of Language*, Oxford: Oxford University Press, pp. 47–68.

Antony, Louise M. (2004), "A Naturalized Approach to the A Priori," *Philosophical Issues*, 14:1–17.

Antony L., and Hornstein, N. (2003), *Chomsky and His Critics*, Oxford: Blackwell.

Armstrong, D. (1968), *A Materialist Theory of the Mind*, London: Routledge and Kegan-Paul.

Baillargeon, R., Scott, R. M., and He, Z. (2010), "False-Belief Understanding in Infants," *Trends in Cognitive Sciences*, 14: 110–18.

Banaji, M. R., and Greenwald, A. G. (2013), *Blindspot: Hidden Biases of Good People*, New York: Delacorte Press.

Baron-Cohen, S. (1996), *Mindblindness: An Essay on Autism and Theory of Mind*, Cambridge: MIT Press.
Barsalou, L. (2010), "Grounded Cognition: Past, Present, and Future," *Topics in Cognitive Science*, 2:716–24.
Bayne, T. (2010), *The Unity of Consciousness*, Oxford: Oxford University Press.
Berker, S. (2009), "The Normative Insignificance of Neuroscience," *Philosophy and Public Affairs* 37(4): 293–329.
Bhalla, M., and Proffitt, D. (1999), "Visual-Motor Recalibration in Geographical Slant Perception," *Journal of Experimental Psychology, Human Perception and Performance* 25(4): 1076–96.
Block, N. (1978/80), "Troubles with Functionalism," in Block, N., *Readings in the Philosophy of Psychology*, vol. 1, Cambridge: Harvard University Press, pp. 261–325.
Block, N. (1995), "On a Confusion about a Function of Consciousness," *Behavioral and Brain Sciences* 18(2): 227–47.
Block, N. (2002a), "Review of Alva Noë," *Action in Perception*, *The Journal of Philosophy*, CII, 5:259–72.
Block, N. (2002b), "The Harder Problem of Consciousness?" *The Journal of Philosophy*, Vol. XCIX, 8:391–425.
Block, N. (2011), "Perceptual Consciousness Overflows Cognitive Access," *Trends in Cognitive Sciences*, 15(12): 567–75.
Brasil-Neto, J., Pascuel-Leone, A. Valls-Solé, J., Cohen, L., and Hallett, M. (1992), "Focal Transcranial Magnetic Stimulation and Response Bias in a Forced Choice Task," *Journal of Neurology, Neurosurgery, and Psychiatry*," 55: 964–66.
Brewer, W. F. (1974)., "There is No Convincing Evidence for Operant or Classical Conditioning in Adult Humans," in W. B. Weimer and D. S. Palermo (eds.), *Cognition and the Symbolic Processes*, Hillsdale, NJ: Erlbaum, pp. 1–42.
Brosnan, S. and de Waal, F. (2003), "Monkeys Reject Unequal Pay," *Nature*, 425: 297–99.
Brown, R. (1973) *A First Language: the Early Stages*, Cambridge: Harvard University Press.
Bruner, J. (1957), "On Perceptual Readiness," *Psychological Review*, 64: 123–52.
Burge, T. (2010), *Origins of Objectivity*, Oxford: Oxford University Press.
Carey, S. (2009), *The Origin of Concepts*, Oxford: Oxford University Press.
Carnap, R. (1947/56), "Meaning Postulates," in his *Meaning and Necessity*, Chicago: University of Chicago Press, pp. 222–9.
Carruthers, P. (2006), *The Architecture of the Mind*, Oxford: Oxford University Press.
Carruthers, P. (2011), *The Opacity of Mind*, Oxford: Oxford University Press.
Chalmers, D. (1996), *The Conscious Mind*, Oxford: Oxford University Press.
Chisholm, R. (1957), *Perceiving: a Philosophical Study*, Ithaca: Cornell University Press.
Chomsky, N. (1959), "Reviews: Verbal Behavior by B. F. Skinner," *Language* 35(1): 26–58.
Chomsky, N. (1965), *Aspects of the Theory of Syntax*, Cambridge: MIT.
Chomsky, N. (1968/72), *Language and Mind*, New York: Harcourt, Brace and World.
Chomsky, N. (1980), *Rules and Representations*, New York: Columbia University Press.
Chomsky, N. (1981), *Lectures on Government and Binding*, Dordrecht: Foris.
Chomsky, N. (2000), *New Horizons in the Study of Language*, Cambridge, UK: Cambridge University Press.
Clark, A. (2011), *Supersizing the Mind*, Oxford and New York: Oxford University Press.

Clayton, N, Emery, N., and Dickenson, A. (2006), "The Rationality of Animal Memory: Complex Caching Strategies of Western Scrub Jays," in S. Hurley, and M. Nudds (eds.), *Rational Animals*, Oxford: Oxford University Press, pp. 197–216.
Collins, J. (2008), "Knowledge of Language Redux," *Croatian Journal of Philosophy,* 7: 3–42.
Collins, J. (2009), "The Perils of Content," *Croatian Journal of Philosophy,* 9: 259–89.
Cosmides, L. and Tooby, J. (1992). "Cognitive Adaptations for Social Exchange," in J. Barkow, L. Cosmides, and J. Tooby (eds.), *The Adapted Mind*, Oxford: Oxford University Press, pp. 163–228.
Cowie, F. (1999), *What's Within: Nativism Reconsidered*, Oxford: Oxford University Press.
Crain, S. and Thornton, R. (1994/2006), "Acquisition of Syntax and Semantics," in M. Traxlwer, and M. Gernsbacher (eds.), *Handbook of Psycholinguistics*, 2nd ed., London: Academic Press, pp. 1073–110.
Cummins, R. (1996), *Representation*, Cambridge: MIT Press.
Davidson, D. (1963/80), "Actions, Reasons and Causes," in his *Essays on Actions and Events*, Oxford: Oxford University Press, pp. 3–19.
Davidson, D. (2001), *Subjective, Intersubjective, Objective*, Oxford: Clarendon Press, 2001.
Dennett, D. (1987), *The Intentional Stance*, Cambridge: MIT Press.
Dennett, D. (1991), *Consciousness Explained*, Boston: Little, Brown & Co.
Devitt, M. (1996), *Coming to Our Senses*, Cambridge: Cambridge University Press.
Devitt, M. (2006), *Ignorance of Language*, Oxford: Oxford University Press.
Devitt, M. (2013), "Intuitions are not the Voice of Competence," in M. Haug (ed.), *Philosophical Methodology: The Armchair or the Laboratory?* London: Routledge, pp. 268–93.
Dilthey, W. (1927/76), "The Understanding of Other Persons and their Life-Expressions," in P. Gardiner (ed,), *Theories of History.* Trans. J. Knehl. Glencoe, IL: Free Press, 1959, pp. 213–25.
Doris, J. (2002), *Lack of Character: Personality and Moral Behavior.* New York: Cambridge University Press.
Dretske, F. (1981), *Knowledge and the Flow of Information*, Cambridge: MIT Press.
Dretske, F. (1988), *Explaining Behavior: Reasons in a World of Causes*, Cambridge: MIT Press.
Durkheim, E. (1897/1951), (1897) [1951]. *Suicide: A Study in Sociology*, New York: The Free Press.
Dwyer, S. (2009), "Moral Dumbfounding and the Linguistic Analogy: Implications for the Study of Moral Judgment," *Mind & Language,* 24: 274–96.
Ericsson K. and Simon H. (1993), *Protocol Analysis: Verbal Reports as Data*, Cambridge: MIT Press.
Erwin, E. (1996), *A Final Accounting: Philosophical and Empirical Issues in Freudian Psychology*, Cambridge: MIT Press.
Fausto-Sterling, A. (1997). "Beyond Difference: A Biologist's Perspective," *Journal of Social Issues*, Vol. 53(2): 233–58.
Fechner, G. T. (1860/1965), "Elements of Psychophysics," in R. Herrnstein, and E. Boring (eds.), *A Source Book in the History of Psychology*, Cambridge, MA: Harvard University Press, pp. 66–75.
Firestone, C. and B. Scholl (2014), "Top-Down Effects Where None Should Be Found: The El Greco Fallacy in Perception Research," *Psychological Science*, 25(1): 38–46.
Fodor, J. (1975), *The Language of Thought*, New York: Crowell.
Fodor, J. (1983), *Modularity of Mind*, Cambridge: MIT Press.

Fodor, J. (1987), *Psychosemantics*, Cambridge: MIT Press.
Fodor, J. (1991), *A Theory of Content and Other Essays*, Cambridge: MIT Press.
Fodor, J. (1998), *Concepts: Where Cognitive Science Went Wrong*, Oxford: Oxford University Press.
Fodor, J. (2000), *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*, Cambridge: MIT Press.
Fodor, J. and Lepore, E. (2010), "Why Meaning (Probably) Isn't Conceptual Role," in D. Byrne, and M. Kölbel (eds.), *Arguing About Language*, London: Routledge, pp. 15–35.
Fodor, J. and Pylyshyn, Z. (1988), "Connectionism and Cognitive Architecture", *Cognition*, Vol. 28(1–2): 3–71.
Foot, P. (1967) "The Problem of Abortion and the Doctrine of the Double Effect," *Oxford Review*, 5:5–15.
Foucault, M. (1961/2006). *Madness and Civilization: A History of Insanity in the Age of Reason*, ed. Jean Khalfa, trans. Jonathan Murphy and Jean Khalfa London: Routledge.
Freud, S. (1916/89), *Introductory Lectures on Psychoanalysis* (trans. by J. Strachey), New York: Norton (Liveright) Publishing Co.
Frith, U. (1989), *Autism: Explaining the Enigma,* Oxford: Blackwell Publishers.
Gallistel, C. (1990), *The Organization of Learning*, Cambridge, MA: Bradford Books/ MIT Press.
Gazzaniga, M. (2000), "Cerebral Specialization and Interspheric Communication. Does the Corpus Collosum Enable the Human Condition?" *Brain*, 123: 1293–336.
Gazzaniga, M., Bogen, J. and Sperry, R. (1962), "Some Functional Effects of Sectioning the Cerebral Commissures in Man," *Proceedings of National Academy of Science,* 48(10): 1765–9.
Gendler, T. (2008), "Alief and Belief in Action (and Reaction)," *Journal of Philosophy,* 105(10): 634–63.
Gilligan, C. (1982), *In a Different* Voice, Cambridge: Harvard University Press.
Geldard, F. and Sherrick, C. (1978), "The Cutaneous 'Rabbit': A Perceptual Illusion," *Science*, 178 (4057): 178–9.
Gigerenzer, G. (2001), "The Adaptive Mind," in G. Gigerenzer, and R. Seltin (eds.), *Bounded Rationality: the Adaptive Toolkit*, Cambridge: MIT Press, pp. 37–50.
Goldman, A. (1986), *Epistemology and Cognition*, Harvard University Press.
Goldman, A. (2006), *Simulating Minds*, Oxford: Oxford University Press.
Gopnik, A. (1993), "How We Know Our Own Minds: The Illusion of First-Person Knowledge of Intentionality," *Behavioral and Brain Sciences*, 16: 1–14.
Gordon, R. (2001), "Simulation and Reason Explanation: The Radical View," *Philosophical Topics,* 29(1–2): 175–92.
Greene, J. (2013), *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*, London: Penguin Press.
Grunbaum, A. (1984), *The Foundations of Psychoanalysis: A Philosophical Critique*, Berkeley: University of California Press.
Hacking, I. (1998), *Rewriting the Soul: Multiple Personality and the Sciences of Memory*, Princeton, NJ: Princeton University Press.
Haidt, J. (2001), "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment," *Psychological Review,* 108(4): 814–34.
Hale, B. and Wright, C., 2000, "Implicit Definition and the A Priori," in P. Boghossian, and C. Peacocke (eds.), *New Essays on the* A Priori, Oxford: Clarendon Press, pp. 286–319.

Hanson, N. (1958/2010), *Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science*, Cambridge: Cambridge University Press.
Harman, G. (1999), "Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error," *Proceedings of the Aristotelian Society,* 99: 315–31.
Horgan, T. and Tiensen, J. (1994), "A Non-Classical Framework for Cognitive Science," *Synthese,* 101(3): 305–45.
Horwich, P. (1998), *Meaning*, Oxford: Oxford University Press.
Horwich, P. (2012), *Wittgenstein's Meta-Philosophy*, Oxford: Oxford University Press.
Hursthouse, R. (1999), *On Virtue Ethics*. Oxford and New York: Oxford University Press.
Isen, A. M. and Levin, P. F., (1972), "Effect of Feeling Good on Helping: Cookies and Kindness," *Journal of Personality and Social Psychology,* 21: 384–8.
Jackendoff, R. (1987), *Consciousness and the Computational Mind*, Cambridge: MIT Press.
Jackson, F. (1982), "Epiphenomenal Qualia," *The Philosophical Quarterly*, 32: 127–37.
Jackson, F. 1998, *From Metaphysics to Ethics*, Oxford: Clarendon Press.
James, W. (1890/1981), *Principles of Psychology*, (two vols), Cambridge: Harvard University Press.
Kahneman, D. (2012) *Thinking, Fast and Slow*, London: Penguin.
Kamtekar, R. (2004), "Situationism and Virtue Ethics on the Content of Our Character," *Ethics,* 114: 458–91.
Kentridge, Robert W., Heywood, Charles A., and Weiskrantz, Lawrence (2004), "Spatial Attention Speeds Discrimination Without Awareness in Blindsight," *Neuropsychologia*, 42(6): 831–5.
Kim, J. (1994), "What is 'Naturalized Epistemology'?"in H. Kornblith (ed.), *Naturalizing Epistemology*, Cambridge: MIT Press, a Bradford Book, pp. 33–56.
Kolers, P. and von Grünau, M. (1976), "Shape and Color in Apparent Motion," *Vision Research*, 16: 329–95.
Kornblith, H. (1994), "Introduction: What is Naturalistic Epistemology," in H. Kornblith (ed.), *Naturalizing Epistemology*, Cambridge: MIT Press, a Bradford Book, pp. 1–14.
Kuhn, T. (1962/2012), *The Structure of Scientific Revolutions*, 50th anniversary edition, Chicago, IL: University of Chicago Press.
Lappin, S. and Shieber, S. (2007), "Machine Learning Theory and Practice as a Source of Insight into Universal Grammar," *Journal of Linguistics*, 43: 1–34.
Lashley, K. (1951), "The Problem of Serial Order in Behavior" in L. Jeffress (ed.), *Cerebral Mechanisms in Behavior*, New York: Wiley, pp. 112–36.
Leslie, A. (1987), "Pretence and Representation: The Origins of Theory of Mind," *Psychological Review*, 94: 412–26.
Leslie, S. (forthcoming) "The Original Sin of Cognition: Fear, Prejudice, and Generalization," *The Journal of Philosophy.*
Levine, J. (1983), "Materialism and Qualia: The Explanatory Gap," *Pacific Philosophical Quarterly*, 64: 354–61.
Levine, M. (2002) (ed.), *Analytic Freud: Philosophy and Psychoanalysis*, London: Routledge.
Lewis, D. (1972), "Psychophysical and Theoretical Identities," *Australasian Journal of Philosophy* 50(3): 249–58.
Libet, B. (2004), *Mind Time: The Temporal Factor in Consciousness*, Cambridge: Harvard University Press.
Lycan, W. (1996), *Consciousness and Experience*, Cambridge: MIT.
McClelland, J., Rumelhart, D., et al. (1986), *Parallel Distributed Processing*, Volume II, Cambridge: MIT Press.

McDowell, J. (1994), *Mind and World*, Cambridge, MA: Harvard University Press.
McLaughlin, B. and Warfield, T. (1994), "The Allure of Connectionism Reexamined," *Synthese* 101: 365–400.
Marcus, G. (2001), *The Algebraic Mind: Integrating Connectionism and Cognitive Science*, Cambridge: MIT Press, a Bradford Book.
Marks, C. (1981), *Commissurotomy, Consciousness and the Unity of Mind*, Cambridge: MIT Press.
Marr, D. (1982), *Vision*, San Francisco: Freeman & Co.
Mathews, K. E. and Cannon, L. K. (1975), "Environmental Noise Level as a Determinant of Helping Behavior," *Journal of Personality and Social Psychology*, 32: 571–7.
Mikhail, J. (2011), *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*, New York: Cambridge University Press.
Milgram, S. (1974), *Obedience to Authority*. New York: Harper and Row.
Millikan, R. (1984), *Language and Other Biological Categories*, Cambridge: MIT Press.
Millikan, R. (1995), "Pushmi-Pullyu Representations," in J. Tamberberlin (ed.), *Philosophical Perspectives*, Vol. 9 (AI, Connectionism and Philosophical Psychology), Atascadero, CA: Ridgeview Publishing Company, pp. 185–200.
Murphy, D. (2006), *Psychiatry in the Scientific Image*, Cambridge: MIT Press.
Milner, D. and Goodale, M. (1995), *The Visual Brain in Action*, Oxford: Oxford University Press.
Nagel, T. (1971/79), "Brain Bisection and the Unity of Consciousness," in his *Mortal Questions*, Cambridge: Cambridge University Press, pp. 147–64.
Nagel, T. (1974/79), "What It's Like to be a Bat," in his *Mortal Questions*, Cambridge: Cambridge University Press, pp. 165–80.
Neander, K. (2012), "Toward an Informational Teleosemantics," in J. Kingsbury and D. Ryder (eds.), *Millikan and Her Critics*, Wiley Blackwell, pp. 21–41.
Neander, K. (forthcoming), *The Emergence of Content: Naturalizing the Representational Power of the Mind*, Cambridge: MIT Press).
Newell, A. and Simon, H. (1972), *Human Problem Solving*, Englewood Cliffs, NJ: Prentice Hall.
Nichols, S. and Stich, S. (2003), *Mindreading*, Oxford: Oxford University Press.
Nisbett, R. and Wilson, T. (1977), "On Telling More Than We Can Know," *Psychological Review*, 84(3): 231–59.
O'Regan, J. K. and Noë, A. (2001), "A Sensorimotor Account of Vision and Visual Consciousness," *Behavioral and Brain Sciences* 24(5): 939–73.
Papineau, D. (1987), *Reality and Representation*, Oxford: Blackwell.
Parfit, D. (1984), *Reasons and Persons*, Oxford: Oxford University Press.
Perry, G. (2012), *Behind the Shock Machine*, New York: the New Press.
Pietroski, P. (2010), "Concepts, Meanings, and Truth: First Nature, Second Nature, and Hard Work," *Mind and Language*, 25: 247–78.
Pinker, S. (2003), *The Blank Slate: The Modern Denial of Human Nature*, New York: Viking Penguin.
Preston, J. and Bishop, M. (2002), *Views into the Chinese Room*, Oxford: Oxford University Press, pp. 201–25.
Prinz, J. (2002), *Furnishing the Mind: Concepts and their Pereptual Basis*, Cambridge MIT Press.

Prinz, J. (2002), "A Neurofunctional Theory of Consciousness," in A. Brook, and K. Aikins (eds.), *Cognition and the Brain: the Philosophy and Neuroscience Movement*, Cambridge University Press, pp. 381–96.
Prinz, J. (2012), *The Conscious Brain*, Cambridge: MIT Press.
Putnam, H. (1960/75), "Minds and Machines," in H. Putnam (ed.), *Philosophical Papers*, vol. 2, Cambridge University Press, pp. 362–85.
Putnam, H. (1965/75), "The Analytic and the Synthetic," in *Philosophical Papers*, vol. 2, Cambridge University Press, pp. 33–69.
Pylyshyn, Z. (1984), *Computation and Cognition*, Cambridge: MIT Press.
Quine, W. (1953a/61), "On What There Is," in his *From a Logical Point of View and Other Essays*, New York: Harper and Row, pp. 1–19.
Quine, W. (1953b/61), "Two Dogmas of Empiricism," in his *From a Logical Point of View and Other Essays*, New York: Harper and Row, pp 20–46.
Quine, W. (1969), "Epistemology Naturalized," his *Ontological Relativity and Other Essays*, New York: Columbia University Press, pp. 69–90.
Rawls, J. (1971), *A Theory of Justice*, Cambridge: Harvard University Press.
Rey, G. (1997a), *Contemporary Philosophy of Mind: A Contentiously Classical Approach*, Oxford: Blackwell.
Rey, G. (1997b), "A Question about Consciousness," (with postscript), in N. Block, O. Flanagan, and G. Gülzedere, *The Nature of Consciousness: Philosophical Debates*, Cambridge: MIT Press, pp. 461–82.
Rey, G. (2009), "Concepts, Defaults, and Internal Asymmetric Dependencies: Distillations of Fodor and Horwich," in N. Kompa, C. Nimtz, and C. Suhm (eds.), The A Priori and *Its Role in Philosophy*, Paderborn: Mentis, pp. 185–204.
Rey, G. (2013a), "The Analytic-Synthetic Distinction," *Stanford Encyclopedia of Philosophy,* <http://plato.stanford.edu/entries/analytic-synthetic> (available only online). Accessed October 28, 2015.
Rey, G. (2013b), "The Possibility of a Naturalistic Cartesianism Regarding Intuitions and Introspection," in M. Haug (ed.), *Philosophical Methodology: The Armchair or the Laboratory?* London: Routledge, pp. 243–67.
Rorty, R. (1991), *Essays on Heidegger and Others*, Cambridge: Cambridge University Press.
Rosch, E. (1978/99), "Principles of Categorization," in E. Margolis, and S. Lawrence (eds.), *Concepts: Basic Readings*, Cambridge: MIT Press, pp. 189–206.
Rosenthal, D. (1993), "Thinking That One Thinks," in M. Davies and R. W. Humphreys (eds.), *Consciousness: Psychology and Philosophical Essays*, Oxford: Blackwell, pp. 198–223.
Ruddick, S. (2002), *Maternal Thinking: Toward a New Politics of Peace*, Boston, Ma.: Beacon Press.
Rupert, R. (2009), *Cognitive Systems and the Extended Mind*. New York: Oxford University Press.
Ryle, G. (1949), *The Concept of Mind*, London: Hutcheson.
Saunders, K. (2014), "Investigating the Psychological Foundations of Moral Judgment," Ph.D. dissertation, Dept of Psychology, Rutgers University.
Schechter, E. (forthcoming), "The Subject in Neuropsychology: Individuating Minds in the Split-Brain Case," *Mind and Language*.
Schütze, Carson (1996), *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*, Chicago: University of Chicago Press.

Schwitzgebel, E. (2014), "Introspection," in *Stanford Encyclopedia of Philosophy*, <http://plato.stanford.edu/entries/introspection/>. Accessed September 25, 2015.
Searle, J. (1992), *The Rediscovery of the Mind*, Cambridge: MIT Press, A Bradford Book.
Shoemaker, S. (1981), "Varieties of Functionalism," *Philosophical Topics,* 12(1): 93–119.
Shoemaker, S. (ed.) (1996), *The First-Person Perspective and Other Essays*, New York: Cambridge University Press.
Siegel, S. (2010), *The Contents of Visual Experience*. New York: Oxford University Press.
Sinnott-Armstrong, W. (2008), *Moral Psychology,* Vol. 1: *The Evolution of Morality;* Vol. 2: *The Cognitive Science of Morality*, Vol. 3: *The Neuroscience of Morality*, Cambridge, MA: MIT Press.
Skinner, B. (1938), *The Behavior of Organisms: An Experimental Analysis*, New York: Appleton-Century.
Smith, E. and Medin, D. (1981), *Categories and Concepts*, Cambridge: Harvard University Press.
Smolensky, P. (1988), "On the Proper Treatment of Connecitonism," *Behavioral and Brain Sciences*, 11: 1–23.
Sosa, E. (1983), "Nature Unmirrored, Epistemology Naturalized," *Synthese* 55(1): 49–72.
Spelke, E. S. (1998), "Nativism, Empiricism, and the Origins of Knowledge," *Infant Behavior and Development*, 21: 181–200.
Sperling, G. (1960), "The Information Available in Brief Visual Presentations," *Psychological Monographs,* 74: 498 (whole issue).
Stampe, D. (1977), "Towards a Causal Theory of Linguistic Representation," *Midwest Studies in Philosophy*, Minneapolis: University of Minnesota Press, pp. 42–63.
Steele, C., (1997), "A Threat in the Air: How Stereotypes Shape Intellectual Identity and Performance," *American Psychologist*, 52(6): 613–29.
Stern, D. (1985), *The Interpersonal World of the Infant*, New York: Basic Books.
Thorndike, E. (1911), *Animal Intelligence: Experimental Studies*, New York: Macmillan.
Titchener, E. (1909), *Lectures on the Experimental Psychology of the Thought-Processes,* New York: Macmillan.
van Gelder, T. and Niklasson, L. (1994), "Classicism and Cognitive Architecture," *Proceedings of the Sixth Annual Conference of the Cognitive Science Society*, Hillsdale: Erlbaum, pp. 905–9.
Walker, L. J. (1984), "Sex Differences in the Development of Moral Reasoning: A Critical Review," *Child Development*, 55: 677–91.
Walton, G. (2014), "The New Science of Wise Interventions," *Current Directions in Psychological Science*, 23(1): 73–82.
Wason, P. C. (1966). "Reasoning," in B. M. Foss, (ed.), *New Horizons in Psychology*, Harmondsworth: Penguin, pp. 135–51.
Watson, J. (1913), "Psychology as the Behaviorist Views It," *Psychological Review*, 20(2):158–77.
Weber, M. (1904/49), "Objectivity in Social Science and Social Policy," in E. A. Shils and H. A. Finch (ed. and trans.), *The Methodology of the Social Sciences*, New York: Free Press, pp. 50–112.
Wegner, D. (2002), *The Illusion of Conscious Will*, Cambridge: MIT Press.
Weiskrantz, L. (1997), *Consciousness Lost and Found: A Neuropsychological Exploration*, Oxford University Press.

Wheeler, M. (2011), "Embodied Cognition and the Extended Mind," in J. Garvey (ed.). *The Continuum Companion to Philosophy of Mind, Bloomsbury Companions*, London: Continuum, pp. 220–38.
Wilson, T. (2002), *Strangers to Ourselves*, Cambridge: Harvard University Press.
Wittgenstein, (1953), *Philosophical Investigations*, New York: Macmillan.
Wundt, W. (1892/1901), *Lectures on Human and Animal Psychology*, 2nd. edn. J. G. Creighton, E. B. Titchener, trans., London: Swan Sonnenschein.

CHAPTER 29

NEUROSCIENCE

ADINA L. ROSKIES

1. What is Neuroscience?

Neuroscience is an interdisciplinary study of brains and nervous systems. It is interdisciplinary because it draws from numerous independently established domain-focused disciplines that contribute essential methods and insights to our understanding of the brain, including but not limited to anatomy, physiology, molecular biology, genetics, computational modeling, neuropsychiatry, psychology, and ecology. This diversity confers great empirical power and breadth in investigating the nervous system, and allows researchers to triangulate on questions from a variety of empirical directions. While many neuroscientists primarily study the human brain and nervous system, the work of a sizeable contingent focuses on other species, or on evolutionary relationships. And even though many neuroscientists study nonhuman animals as model systems for understanding human brains, quite a few neuroscientists are interested in the workings of simple neural systems for their own sake.

2. The Relation Between Philosophy and Neuroscience

Philosophy and neuroscience can each influence the other. There is a clear connection between philosophy and neuroscience in philosophy's role as a methodological critic and gadfly. Neuroscience is an interdisciplinary, multifaceted, and constantly evolving field. Every day brings new insights into brain function, and a steady development of novel technologies and methods provides ever more powerful means of interrogating the brain. Philosophy, in its role as disciplinary critic, can contribute to clarification, improvement, and a deeper understanding of these methods and advances (see, for example (Craver, 2007). Philosophy of neuroscience is the branch of philosophy of science that aims to provide a critical analysis of the logic of the methods of neuroscience. Other philosophical

areas focus on other aspects of neuroscience research. Neuroethics, for instance, addresses ethical issues raised by neuroscience practices, as well as the implications of our growing understanding of the brain bases of ethical reasoning, judgment, and motivation.

However, neuroscience can also contribute to philosophical theorizing. Many philosophical problems are concerned with understanding mind and mental phenomena. It is widely accepted that the brain realizes, causes, or otherwise gives rise to mind, and in these areas there is a natural overlap between philosophy and neuroscience. The term "neurophilosophy" refers to the use of insights and data from neuroscience in approaching philosophical questions about mind, including metaphysical questions about the relation of mind and brain, questions about the nature of mental representation and content, consciousness, and even moral theory. The particular issue I address here is whether, and in what ways, neuroscience can illuminate these neurophilosophical questions.

3. The Puzzle That Is the Brain

In order to understand how neuroscience can contribute to philosophical issues, it behooves us to bear in mind the puzzle that is the brain. Under ordinary viewing conditions, a brain looks like a mass of relatively homogeneous pinkish-grey tissue. That brains are composed of discrete elements rather than a continuous matrix of material was a hypothesis and its confirmation a scientific discovery (The Neuron Doctrine (Guillery, 2005)) that relied upon the development of appropriate techniques for visualizing brain anatomy. Needless to say, historically there were no methods for visualizing the structure of brains in living humans: brains had to be removed from the skull post-mortem, and sectioned. When sliced it is visually apparent that there are structures that can be discerned even with the naked eye, such as the folded thin layer of grayish tissue that covers the outside of the cerebral hemispheres (the cortex), the highly articulated lobes of the cerebellum located beneath the cortex, the white tracts of the corpus callosum and the internal capsule, and the subcortical nuclei. However complex this gross neuroanatomy is, it provides no inkling of the incredible complexity of the brain's microstructure: the brain is composed of approximately 10^{12} neurons, each of which makes somewhere between 1,000 and 10,000 connections (synapses) with other neurons. Connecting these neurons are microscopic filaments, the axons and dendritic processes that extend from the cell body and carry information to and from other neurons. One significant challenge in neuroscience lies in figuring out how to identify individual elements of microscopic size and their structural relations in material that looks to the naked eye undifferentiated. This is, in practice, an extremely difficult problem. Given the magnitude of this problem, it is not surprising that so much remains unknown about the fine structure of the human brain.

The ultimate problem that neuroscientists want to solve, of course, is how the brain functions. Here there are also significant technical problems. It is virtually impossible to study function in dead tissue (not entirely impossible because one can study function by examining brain structures for which we have functional correlates), but it is also not trivial to measure function in living tissue, since such measurements are almost always invasive, difficult to target, and difficult to interpret. Almost never does learning about the brain merely

involve unmediated looking and straightforward inference. In short, the study of the brain is one extended example of puzzle-solving, one that involves the development and application of tools, logical and statistical inference, theory construction and testing, and more.

To better orient the reader I briefly mention a few of the underlying theoretical commitments and open questions in approaching the study of the brain, and then give an overview of neuroscientific methods. Perhaps the most fundamental commitment is a commitment to the view that the operation of the brain is just the operation of physical processes governed by natural laws. A second commitment to which almost all subscribe is that behavior and cognition are generated by the operation of brain processes. It is a further question whether all mental events and processes (such as consciousness) have a purely physicalist explanation—one could be a neuroscientist yet have a dualistic view of mentality, as did Nobel Prize winning neuroscientist John Eccles. However, neuroscientists with these views are increasingly in the minority.

In addition to these philosophical commitments are some scientific ones. A common grounding commitment of neuroscience is that structure constrains function. This view upholds the importance of understanding neuroanatomy for understanding brain function. Specificity in brain structure and function has not always been apparent. Largely on the basis of being unable to locate the "engram" (memory), Lashley and Pribram argued that the brain was an equipotential machine, with all parts participating in all functions, and that function could not be localized (Pribram, 1982). However, subsequent evidence from both neuroanatomy and neuropsychology seemed to point to a high degree of modularity in brain structure and function (Fodor, 1983; Treisman and Kanwisher, 1998; Ullman et al., 1993). Current evidence suggests that massive modularity is too strict a characterization of the brain's structure/function organization. One of the primary goals of neuroimaging is to better describe the brain's structure/function relationships. So although the dictum that structure constrains function is a truism, how strong those constraints are and how they operate remains to be determined.

Finally, there are a variety of views about what the relevant aspects of neural function are for understanding various behavioral processes. In other words, what is the "neural code"? Here there are a variety of competing possibilities, such as neural firing rate (how frequently, on average, a neuron fires), the temporal pattern of firing, temporal coordination and/or oscillation between neurons (phase-locked firing), or even features of the low-level physical properties of neuronal function such as chemical fluxes. Although there are arguments and evidence supporting different views, and the verdict is still out, one is forced to make assumptions about the relevant phenomena when designing and executing experiments. These assumptions naturally influence the interpretation of results.

4. Neuroscience Methods: A Brief Introduction

Neuroscience employs a variety of methods for investigating brain structure and function. In what follows I will briefly canvass some of the most prevalent techniques for investigating brain structure and function. Each of these techniques fills an important

niche in terms of its ability to provide information about the spatial and temporal characteristics of neural structure and activity. No single technique can provide all the information we want, and often interpretation of some piece of neuroscience data will depend upon findings from other neuroscientific techniques. Indeed, one indication of the accuracy or reliability of an interpretation is consistency with findings from other techniques.[1]

Among the methods central to neuroscience are the use of stains, antibodies, receptor ligands, or genetic markers to identify cell types, the distribution of receptors throughout the brain, and cellular anatomy. Stains are chemicals applied to neural tissue that has typically been chemically fixed to prevent deterioration, and frozen and sectioned in order to increase the surface area of neural tissue that will come in contact with the stain. The stain will react with or bind to particular elements in the neural tissue in ways that differentiate some elements from the surrounding matrix of tissue in ways that can be visualized with light, fluorescent, or other microscopy. Antibodies bind to specific molecules in the tissue and can be visualized with stains or fluorescent techniques. These methods can provide information about different aspects of neural structure. Tracers, which are markers that are transported toward or away from the cell body, are used to delineate fiber pathways and to determine cellular connectivity. Anterograde tracers are tracers that are injected into a small region of brain tissue, and are taken up by neurons and transported down the axon to the neuron terminals. After transport, the tissue is analyzed, and thus one can see where the neurons with cell bodies at the injection site project. Retrograde tracers, on the other hand, are taken up by axon terminals, and then carried back to the cell body by axonal transport processes. After transport, the tissue can be analyzed to see where the cell bodies are that send their axons to the injection site. Many tracers can only reveal information about the cells they are taken up by: they tell you where those cell project to or from. However, a few tracers work trans-synaptically: they can be transported across a synapse, taken up by the postsynaptic cell, and transported to the cell body of that cell (sometimes with more than one iteration). These sorts of tracers can be used to trace consecutive elements in a neural circuit. Care must be taken with these methods because faulty inferences can occur if nearby fibers (fibers of passage) are damaged by the injections and take up the tracer. Other methods, such as high-power microscopy of various sorts are used to examine the internal structure of neurons and the synapses they make with other neurons. All these are standard techniques employed by neuroanatomists to describe brain structure.

Structural information, although essential, does not alone reveal how the brain works. For that, we have to investigate function. The oldest methods of exploring neural function depended upon the analysis of functional deficits of people or animals in which regions of the brain had been lesioned. But how do we study normal brain operation? Over the past century and a half, various electrophysiological methods have been developed that measure electrical properties of neurons. Neuronal firing involves the generation and propagation of electrical signals, and communication between neurons is

[1] For further information on other techniques mentioned in the figure, see for example (Gazzaniga, Ivry, and Mangun, 2008; Kandel, Schwartz, Jessell, Siegelbaum, and Hudspeth, 2012)..

mediated by chemical neurotransmitters, which are released from the presynaptic neuron and induce electrical signals in the postsynaptic neuron. These electrical signals associated with neural activity can be measured by recording from neural tissue with electrodes. In the laboratory it is often possible to penetrate single neurons and record the electrical changes with intercellular electrodes. This is usually done in slices of brain tissue, kept viable for a time by immersion in an oxygenated bath of fluid with nutrients. In anesthetized and sometimes in awake animals, it is possible to record from electrodes in the brain placed close enough to a single neuron so that the electrical signals from that neuron can be isolated. Such recording can also sometimes be done in humans that are undergoing brain surgery, but opportunities for doing so are rare, and limited by the clinical treatments subjects are undergoing. Finally, recent techniques using multi-electrodes have been developed, allowing researchers to record from numerous (10–100) cells at a time. These cells are all located in a restricted area close to the microelectrode array, and multicellular recording can provide a picture of the activity of both individual and population activity. While many of these electrophysiological techniques are most commonly used in cell culture or in tissue slices, researchers are developing powerful new methods for doing neurophysiology in awake behaving animals. It is rare to have an opportunity to use standard neurophysiological or neuroanatomical techniques in humans, because of the invasiveness of these methods.

The barrier to investigating brain structure and function in humans has been surmounted to some extent with the development of novel brain-imaging methods. These methods allow both anatomical and functional brain data to be acquired in normal, awake (and to some degree, behaving) humans. Magnetic resonance imaging (MRI) and functional MRI (fMRI) enable the acquisition of whole-brain anatomical and functional data respectively. fMRI measures changes in blood flow, volume, and oxygenation, which are correlated with neural activity. Other variations on MRI are increasingly used to map neural connectivity in vivo. However, the temporal and spatial resolution of MRI methods is much coarser than many other techniques for investigating structure and function, and the functional data does not directly reflect neural activity, either spatially or temporally. For these and many other reasons, the interpretation of neuroimages is complicated and fraught. To a large extent the methodological and interpretational difficulties are overlooked by nonscientists who are apt to refer to scientific studies in their arguments, but are familiar only with oversimplified media summaries of the science and tend to be unaware of how far from accurate these often are. Given how prevalent references to neuroimaging are in the philosophical literature, interested readers should consult other sources for more details on the interpretation and pitfalls of neuroimaging (Jones, Buckholtz, Schall, and Marois, 2009; A. L. Roskies & Petersen, 2001) before citing these data in their arguments.

This caveat extends widely: If one intends to use neuroscience information in the service of philosophical argumentation, it would be wise to look beyond popular accounts of neuroscience which tend to obscure or ignore the complexities of interpretation, and gain familiarity with the science itself. Understanding the scope, limitations, and potential confounds of the techniques one appeals to is an essential aspect of using this data (see Figure 29.1).

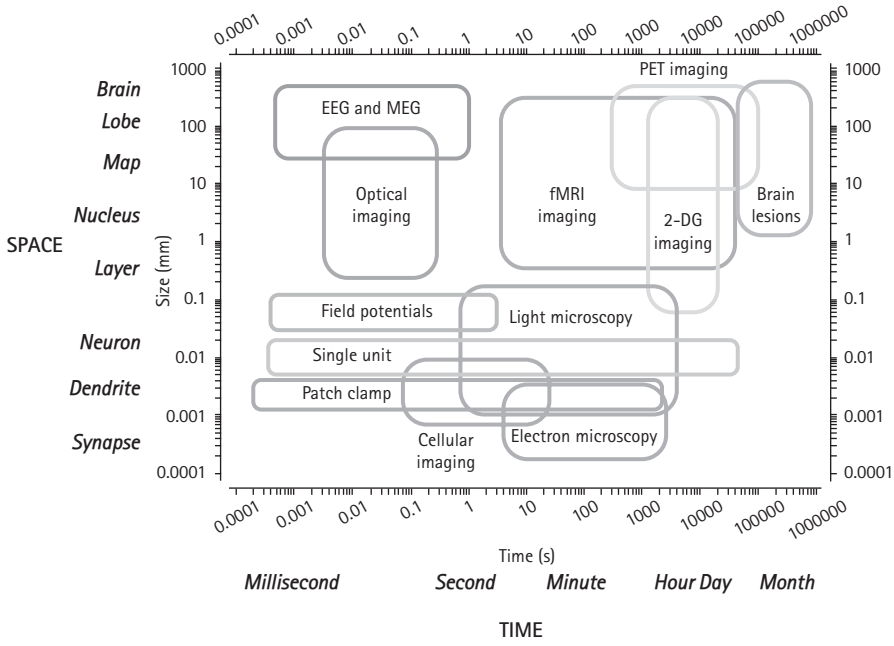


FIGURE 29.1 Graph of spatial and temporal resolution of neuroscientific techniques. Each box represents the spatial scales and temporal discrimination possible with its associated neuroscientific technique. The temporal and spatial scales are graphed logarithmically, ranging from microns to the scale of the entire brain, and from milliseconds to years. Each technique only covers a limited part of the information range. To fully understand brain function we need to integrate information from a variety of neuroscientific techniques, none of which alone can tell us more than a part of the story.

Adapted from Churchland and Sejnowski, 1988

5. Facts and Philosophy

Philosophy is often distinguished from the sciences by its a priori quality, or its lack of reliance on data. This is a misleading characterization, for insofar as philosophy aims to tell us about the world we live in, it is (or should be) as bound by fact as other disciplines. Although philosophical theorizing often seeks to go beyond the actual by addressing modal questions, philosophical theory must still accommodate the facts (or the facts as best we understand them).

Neuroscience provides us facts about a certain segment of our world: the biological world of the brain and nervous system, and the functions that these things support, such as cognition, awareness, emotion, and so on. These facts can often be an important boost to philosophical theorizing. For instance, understanding the facts of split brain patients undoubtedly informed Nagel's discussion of consciousness in Nagel's "Brain Bisection and the Unity of Consciousness." Other data from neuropsychiatry, neuropsychology, and cognitive neuroscience (including psychology) has informed philosophy of mind and

philosophy of psychology. Neuroscience has made some progress toward characterizing the bases of mental representation, and although the work is perhaps too preliminary to bear directly upon the philosophical discourse surrounding the topic (e.g. the language of thought, conceptual and nonconceptual content of experience, theories of semantics), it would be shocking if it did not bear substantively on these debates in the future. The same can be said for work on the neural correlates of consciousness.

In addition to informing or motivating philosophical arguments, there are particular roles that facts from neuroscience can play in philosophical discussions. First, they can potentially be important in providing counterexamples to particular philosophical claims. Second, they can provide existence proofs or proof of possibility. They can also play a role in more complex systematic questions. I will describe each of these in turn, and will describe some examples of how neuroscience has been used (successfully or unsuccessfully) for each. I end with a cautionary tale, which illustrates both how philosophers should be sophisticated consumers of neuroscience data if they plan to make use of them in their arguments, and also how neuroscientists ought to become philosophically astute if they intend to make philosophical arguments using their data. While there is surely much to be gained by hurdling disciplinary boundaries, oversimplification and inattention to detail have led to many a wrong turn.

5.1 Counterexamples

Philosophers often aim to find general principles describing or governing the nature of specific phenomena, such as knowledge, consciousness, or justice. Progress is made when some such principle is discovered, but also when a counterexample is provided that shows that some plausible or accepted principle does not accurately demarcate the boundaries of the phenomenon in question. Typically, counterexamples lead philosophers to either 1) try to revise or precisify their analysis so that it accommodates the proffered counterexample, or 2) reject the counterexample. For example, Gettier cases are counterexamples that most agree demonstrate that knowledge is not equivalent to justified true belief, and consequently much of epistemology since Gettier has sought further conditions on knowledge or precisifications of what counts as justification. Gettier cases and many other counterexamples in philosophy are the result of pure thought experiments, in that they do not rely upon empirical evidence, but rather upon intuitions about cases and abstract reasoning. But thought experiments can be contentious, precisely because people's intuitions about imagined cases often conflict. For example, the possibility of zombies has been offered as a counterexample to physicalism about the mental, but many physicalists are unmoved by zombie arguments, for they claim that zombies are not really possible. One way to counter arguments against possibility is by demonstrating possibility in actuality. Neuroscience and other sciences can potentially contribute in this way, for sometimes, it seems, truth is stranger—and at the very least more credible—than fiction.

5.1.1 *Example: Acquired Sociopathy as a Counterexample to Motive Internalism*

One of the first examples where neuroscience was directly employed to provide a counterexample to a philosophical claim was in Roskies' argument that subjects with acquired

sociopathy (i.e. damage to ventromedial prefrontal cortex in adulthood) are counterexamples to the thesis of motive internalism (or motivational internalism) (Roskies, 2003). That thesis, at least in one of its guises, claims that moral judgments are intrinsically motivating. That is, whenever one makes a moral judgment one is motivated to act in accord with it. This is meant to be not merely a descriptive claim, but a claim of necessity. A number of versions of this thesis have been offered, varying in plausibility (see (Smith, 1993)). Previously, the best counterexample to internalism was the amoralist, a possible but nonactual person who makes moral judgments but is not moved by them (Brink, 1989). The internalist reply to the amoralist was simply that amoralists were impossible. A battle of intuitions ensued. The focus of the debate was changed with the inclusion of real-world examples from neuroscience. Roskies argued that subjects with damage to ventromedial prefrontal cortex make the same moral judgments that normal people make, but do not appear to be motivated by their judgments. These subjects lack physiological responses (skin-conductance responses, or SCRs) that are indications of activation of the autonomic nervous system, a point taken as evidence of lack of motivation. The existence of people with this profile thus suggests that it is possible to make moral judgments without being motivated, and provides resources for arguing that the judgments made are real moral judgments, not just judgments in the "inverted commas" sense.

As with many other counterexamples in philosophy, descriptions of people with ventromedial prefrontal cortical damage did not put the nail in the coffin of motive internalism. Kennett and Fine defended motive internalism by arguing that few people hold to the unrestricted thesis of motive internalism that Roskies attacks, that the absence of an autonomic response is not a measure of lack of motivation, and that these subjects make normal moral judgments (Kennett and Fine, 2007). Indeed, later work with subjects with VM damage suggests that their judgments concur with normals' judgments over a wide range of cases, but that in some cases VM patients tend to make more utilitarian judgments than the majority of normal people (Koenigs et al., 2007). Whether this shows that VM subjects do not make normal moral judgments, or just fail to make normal judgments in a restricted realm remains unresolved (Interestingly, the pattern of the VM subjects' moral judgments aligns with the judgments of some philosophers, calling into question whether equating "the most common" with "normal" is justified). In addition, whether or not the absence of an SCR is adequate evidence for lack of motivation, and what would be considered adequate evidence, has been a matter of continuing debate. Along other lines, Cholbi (Cholbi, 2006) argues that VM subjects do not have moral beliefs, and thus cannot make moral judgments (but see also Roskies, 2006). The debate between motive internalism and externalism continues, but the inclusion of real-world case studies has led to refinement of the philosophical theses at issue, and a shift toward discussion of more concrete phenomena that play a role in our moral economy.

5.2 Providing Alternative Interpretations

In addition to providing potential counterexamples to philosophical claims, by showing how brains work, neuroscience suggests new interpretations of data. One method now standardly used in neuroscience (construed broadly) is computational modeling of neural or neuron-like processes. Modeling can indicate that certain types of functions or

computations are achievable with circumscribed resources. In other words, it can provide proof of principle. It also can aid in the interpretation of function, by showing that candidate interpretations are potentially correct.

5.2.1 *Example: Shape-from-Shading*

One interesting episode from neuroscience comes from the history of connectionism, and illustrates how standard functional interpretation of neuroscientific data can be challenged by a computational model. However, the broader interpretation of this work is that it calls into question our intuitive interpretations of function. Data from neurophysiological recordings of the response properties of cells in the visual cortex provides the basis for assigning function to cell types; the properties to which they respond characterize their "receptive fields". Hubel and Weisel were awarded the Nobel Prize in Medicine in 1981 for their discovery in the 1960s that cells in visual cortex have different and systematic response properties to visual stimuli. They found that some cells in visual cortex respond preferentially to edges in their receptive fields; these cells were identified as "edge detectors". Cells further up in the visual hierarchy were thought to be sensitive to more complex visual features built up of combinations of these simple cells. A taxonomy of these functions was devised that reflected the physiological properties of the cells in early visual cortex. In the award speech at the Nobel ceremony Hubel and Weisel were lauded to have "succeeded in breaking the code of the message which the eyes send to the brain".

The orthodoxy of Hubel and Weisel's functional interpretation remained largely unchallenged for years. However, challenges did arise, and at least one came with the advent of connectionist models in the late 1980s. Lehky and Sejnowski (Lehky and Sejnowski, 1988) built a connectionist model that was trained to distinguish 3D shape from 2D shading information. The model was trained on a variety of smoothly varying convex and concave shapes lit from a particular angle. Through iterative application of a general learning algorithm, the network successfully learned to distinguish convex from concave shapes. Importantly, the network never once was provided with a stimulus that had a sharp edge or contrast line. However, when Lekhy and Sejnowski examined the properties of hidden units of the network (which are the units that largely determine what computation is performed) they found many units that had the activation profiles that matched those of the simple cells in visual cortex that Hubel and Weisel had dubbed edge-detectors!

What is the philosophical take-home message from this episode? It seems there are several possibilities. First, it suggests that our intuitive functional classification of neural computations may not be the correct ones. Perhaps cells in visual cortex should be considered shape-from-shading cells rather than edge detectors—or perhaps not. The modeling shows that neural networks can develop units with certain kinds of response properties when we set the function and thus can be certain of the functional demands the network has to meet, and that the function may not coincide with our intuitive judgments. The shape-from-shading model shows that there are at least two possible legitimate interpretations of the function of cells in visual cortex; there are probably others. This raises the possibility that there is no single correct functional attribution to cells, or that the optimal functional interpretation does not coincide with any of our intuitive judgments (more on this later). So, in the case of simple cells in primary visual cortex, one of the two proffered

interpretations may be correct, or yet another interpretation that does not map easily onto our intuitive functional ontology at all. This last is not unthinkable, since the brain is an organ that needs to conform to a multitude of constraints simultaneously, and varying constraints over time. The study also advances our methodological insight in neuroscience by demonstrating that important functional constraints can be discerned by paying attention to a cell's projective field (characteristics of its output targets) as well as its receptive field. Interestingly, a similar insight emerges from a proper understanding of functionalism, wherein both the input and output of a state are required to characterize its function. In addition, this work points to the potential usefulness of historical approaches in determining function, since it is only by reference to the task the network was designed to solve (and trained on) that its function as a shape-from-shading device becomes clear. This bears some connection to some work on meaning or semantics, such as Millikan's teleosemantic approach (Millikan, 1984), in which meaning depends upon an understanding of proper function, and function is determined by reference to historical/evolutionary context. While the shape-from-shading model probably has more to say to neuroscientists than to philosophers, it provides a salutary caveat about imposing our intuitive conceptual structure on interpretations of function, and this is surely an issue of importance to philosophers.

5.3 Neuroscience and Normativity

In addition to offering candidate counterexamples and existence proofs or proof of possibility, neuroscience may play a role in larger systematic debates in philosophy.

Perhaps the most acrimonious arena of neuroscientific influence on philosophy is in the area of normativity: What role, if any, can neuroscience play in telling us what we ought to do, what is right, what is just? Since Hume, the dictum "you can't derive an ought from an is" has been taken to be a philosophical truism, although more recently, many varieties of naturalism have sought to show that the is-ought gap is not as unbridgeable as was once thought. The latest in this debate comes from neuroscience, regarding the implications of a series of papers by neuroscientist Joshua Greene and criticisms of his work by his Harvard colleague Selim Berker.

5.3.1 *Example: Is Neuroscience Normatively Insignificant?*

Greene's work provides evidence that our moral judgments arise from the interplay of two identifiable brain systems—one involving activation of brain regions mostly associated with emotional processing (including activation of subcortical limbic structures), and the other by the activation of cortical systems more typically associated with rational thought and calculation (Greene, Sommerville, Nystrom, Darley, and Cohen, 2001).[2] Greene characterizes these two systems as leading to different types of moral judgments: what Greene

[2] We should be careful to note that the brain areas involved in emotional and cognitive processing are not mutually exclusive or entirely stable, and perhaps more importantly, that our intuitive classification of tasks as emotional vs. cognitive is highly suspect.

calls "characteristically deontological judgments" and "characteristically consequentialist judgments" respectively (although he is careful to say that they may not map cleanly onto the philosophical landscape)(Greene, 2008). Greene suggests that these dual systems sometimes act in opposition to one another, and that some kinds of moral dilemmas—those that involve personal force—strongly activate the emotional system so that it overrides the rational system, leading to "characteristically deontological judgments"(Greene, 2008; Greene, Nystrom, Engell, Darley, and Cohen, 2004). Thus far, the work mentioned is purely descriptive. However, there is a normative upshot. Greene also argues that the neuroscientific evidence about how our moral judgments work calls into question the validity of our deontological judgments, and that we ought therefore to discount our deontological, but not our consequentialist judgments (Greene, 2008). If Greene is right, we have the surprising result that science can provide us reason to prefer one normative ethical theory over another: a very significant philosophical outcome, and clearly contrary to Hume's dictum.

Berker (Berker, 2009) responds fiercely to a number of Greene's claims, but the arguments that concern us most here are Berker's arguments that neuroscience has no normative significance. In particular, Berker argues that neuroscientific information "does no work" in any of what he deems Greene's "better" arguments, and that all the work is done by appeal to substantive normative intuitions. Berker also calls attention to an almost inevitable problem with empirical work: We rarely have the entire story, and that limits the kinds of conclusions we may be warranted in making. He points out that Greene's characterization of the factors that modulate our judgments is rough and incomplete, and until we can accurately characterize the factors to which these two systems respond, we cannot determine whether it is only the emotional system that responds to morally irrelevant factors: privileging one over the other is premature, and deciding which to privilege is potentially as subject to battles of intuition as are the classic deontological/consequentialist debates. Kamm (Kamm, 2009) echoes these concerns, and argues that deontologists can make judgments that Greene categorizes as characteristically consequentialist, casting doubt upon the identification of normative theory with empirical data that drives Greene's criticism of deontology. She also points out that the difference between an act being a duty and being permissible is important to the debate, but left ambiguous in Greene's experiments.

Returning to our main stalking horse, the primary argument Greene makes for discounting our deontological judgments is that these judgments respond to factors in the dilemmas that are morally irrelevant. He points to the factors that he has identified experimentally as influencing our moral judgments via the emotional subsystem. Berker is correct that the judgment that some factors to which we respond are morally irrelevant stems from substantive moral intuitions that we bring to bear on the analysis. But does this mean that neuroscience does no work in Greene's arguments?

Not necessarily. For his part, Greene freely acknowledges that identifying some factors as morally relevant or irrelevant depends upon substantive moral intuitions; he concurs that his argument does not derive an ought from an is. He writes, "Thus, as Berker argues, any valid normative conclusions reached on the basis of scientific research must also invoke one or more non-scientific normative premises. However, it does not follow from this conclusion that scientific results inevitably do "no work" in such normative arguments." Greene objects that characterizing the debate in terms of the normative significance of

neuroscience is mistaken—it is a debate not about neuroscience (i.e the neuroimaging data) alone, but about the role of psychological data more generally.

Greene points out, rightly, that descriptive or factual premises often are essential to—that is, "do work"—in normative arguments. While he agrees that the neuroscientific details of the system do not matter directly, the fact that there is a dual-process system (a fact not wholly dependent upon neuroscience evidence, but bolstered by it) is essential to the argument. Greene argues that the dual-process theory, and the evidence of which process is at work when, supports the view that one system and not the other responds to morally irrelevant factors. And neuroscience, he argues, supports the existence of dual opponent processes that are entrained by different situational factors.

There is more agreement here than meets the eye. The important point—and one that Greene and Berker agree on—is that neuroscience evidence (and scientific evidence more generally) provides only descriptive facts. However, as Greene argues, those facts can influence and be relevant to our normative judgments, and can play an important role in normative arguments. It seems to be primarily a verbal dispute whether neuroscience "does work" in normative arguments. If one interprets "doing work" as playing a direct justificatory role independently of other normative commitments, as I believe Berker does, then neuroscience does not do work in normative arguments. But if one interprets "doing work" in a more moderate way, such as supporting factual claims that serve as premises in arguments along with other normative premises, as Greene does, then neuroscience certainly does "do work" in such arguments. Either way, neuroscience could be important or even indispensable for making particular arguments with normative import. In the end, Berker does acknowledge that neuroscience can play an indirect role in judgments of normativity, by giving us prima facie reason to question the outputs of certain systems, but not by playing a direct justificatory role. I think this is exactly right. For her part Kamm, while taking issue with Greene's actual experiments, suggests other experiments she thinks would be more illuminating, also implying that empirical work can bear on philosophical questions (see also Kahane, 2012).

Here are two more situations in which normative conclusions are drawn from arguments in which non-normative premises play a crucial role. Suppose, for instance, that with future neuroscience we develop a more objective measure of intelligence than the standard intelligence tests used today, which are criticized as being biased by socioeconomic factors and cultural factors. And suppose that we as a society accept the normative principle that people of substandard intelligence should not be eligible for the death penalty, as the Supreme Court ruled recently in *Atkins v. Virginia*. What sort of penalty should a defendant be faced with? It is not far-fetched to think that whether or not a future convicted defendant is faced with the death penalty will rest on the outcome of neuroscientific tests. It is true that the original validation of the neuroscientific test rests on behavioral or psychological work, but that is no reason to discount the evidential value, and logical role, of neuroscience in this example. Along these lines, as we come to understand more about agency and how different capacities relate to responsibility, and discover the neuroscientific underpinnings of some of these capacities, we may expect neuroscience to play a role in normative arguments regarding culpability.

As another example, we point to the recent discovery that some people previously thought to be in a vegetative state, unresponsive and unaware of their environment, may

not be so (Owen et al., 2006). Brain scans have revealed that some people classified as vegetative are able to respond to instructions to imagine themselves navigating through their house, or playing tennis—two mental states that can be distinguished by neuroimaging. This suggests both awareness and comprehension of instructions, and the volitional ability and executive control to comply. By most standards, these patients are conscious (at least at times), and given our views about the moral implications of consciousness, we have a duty to provide a different level or type of care, and to try to respect their autonomy, and so forth, than we would have were they indeed in a persistent vegetative state.

5.4 Neuroscience and the Foundations of Our Cognitive Ontology

One of the perennial questions of philosophy from Plato on has been how to characterize and subdivide the mind and mental states. From impressions and ideas, sensations and perceptions, concepts and intuitions, and the various states of folk psychology (beliefs, desires, hopes, fears, etc.), philosophers have been trying to provide a principled taxonomy of mental states. This interest goes well beyond the realm of introspectively accessible states to other states and/or processes that may be important to the computational economy of the mind. In general, the differentiation of mental states has been made according to their input and output characteristics, or their function.

Neuroscience is interested in function too: in particular in identifying the function of particular neural states. This raises the prospect that neuroscience could influence and assist in the philosophical project of state identification, or vice versa. We have already seen how neural modeling may alter the way that we identify the functional role of neural building blocks. More recently, novel neuroscientific techniques such as functional MRI are seen as a tool for exploring the functional ontology of the mind.

Here, the naïve hope is that looking at patterns of brain activation (or networks of activated areas) during the performance of certain tasks will lay bare the functional joints of the mind. However, standard methods for analyzing fMRI data depend upon our intuitive taxonomy of functional states in a way that will not clearly tell us whether that intuitive taxonomy is approximately correct or deeply off-base (Anderson, 2015; Poldrack, 2006; Poldrack and Wagner, 2004). Some neuroscientists are cognizant of this, and are trying to develop psychologically theory-neutral, data-driven approaches to delineating function (Yarkoni, Poldrack, Nichols, Van Essen, and Wager, 2011). It remains to be seen whether such approaches will supply us with a different (yet intelligible) functional ontology, and how different it will be. It also remains to be seen whether such approaches are truly theory-neutral with respect to psychological ontology, or are beholden to other theories that may affect interpretation of the data. Finally, we confront the question of whether we are forever bound by the conceptual resources we begin with, or whether we can break out of a mistaken ontology and bootstrap our way to an entirely different theoretical perspective. It is an interesting question whether this issue of the possibility of conceptual change has special bite in the context of psychology, or whether psychology is in the very same position regarding the possibility of radical theory-change as are the other sciences. Regardless of how we end up answering these questions, the goal of understanding

the functional subdivisions of mentality is one which is clearly shared by philosophers and neuroscientists alike, and one for which we can see clear contributions to philosophical theorizing from the neurosciences.

5.5 Cautionary Tales

Thus far I have described ways in which data from neuroscience has been or could be employed in philosophical theorizing in relatively productive ways. But the brain is a complicated organ, and philosophical problems are difficult. Unsurprisingly, bringing the two together is not a trivial undertaking. In this section I describe a number of attempts to use neuroscience to inform philosophical argument about free will. Each of these fails, either because the science is not properly understood, or the philosophical problem is not properly appreciated (or both). I use these examples from the free will literature to illustrate the importance of a sophisticated understanding of both aspects of the problem.

5.5.1 *The Problem of Determinism*

Neuroscience evidence is sometimes invoked as a potential source of empirical answers to metaphysical questions. For example, not a few people believe that neuroscience will challenge belief in free will by showing that the brain is a deterministlc system. For instance, Balaguer (Balaguer, 2004) writes, "the question of whether libertarianism is true just reduces to the question of whether some of our torn decisions are undetermined at the moment of choice. This, of course, is a *straightforward empirical question* about the neural events that are our torn decisions" (my emphasis). Balaguer thus advocates looking at the brain during difficult ("torn") decisions, in which the subject, prior to choosing, has no idea which option she will choose. The naïve idea is that just examining the neural activity will reveal whether that activity is the result of deterministic or indeterministic processes—that is, whether the decision is undetermined at the moment of choice. However, this could not be farther from the truth. There are several reasons why neuroscience will not solve the problem of determinism. First, predictability is the operational handle we have on a system's dynamics if we do not know its governing laws, but predictability is not a good guide to determinism. Chaotic activity, which is likely widespread in the brain, could be the result of either deterministic or indeterministic processes. And predictable activity, within some margin of error (which is the best that our neuroscientific theories can provide) can be the result of indeterministic as well as deterministic processes. Moreover, because of the multiple levels of organization in the brain, and the mindboggling interconnectivity of neurons, we could never have enough empirical information about the brain to enable us to determine whether an unexpected event was due to indeterminism, or to the deterministic evolution of unknown details of other parts of the system. The thought that we can answer the question of determinism by an empirical assault on the brain is fundamentally misguided, and shows an inadequate appreciation of both the theoretical subtleties and the empirical realities of brain research.

5.5.2 *The Timing of Conscious Awareness*

The work of Benjamin Libet is perhaps the most discussed neuroscientific work in philosophy. Libet purported to demonstrate, using neuroscientific methods, that our conscious intentions could not be the initiator of our actions, and thus that we do not in fact have "free will." Instead, unconscious brain processes initiate actions, and we merely misinterpret our intentions as doing causal work. The only hope for rescuing free will, as it is traditionally conceived, Libet thought, is by showing that we have (conscious) veto power over our unconsciously initiated actions. Libet failed to establish that we have this veto ability, what has come to be called "free won't". Libet's results have been widely taken to show that we lack free will, but I think that the philosophical acceptance of this view is based on an insufficient understanding of the underlying neurobiology and the methods employed.

The neuroscientific results that led Libet to this radical claim are the following. 1) He showed that an electrical potential on the scalp (called the "readiness potential" or RP) precedes self-initiated finger movement by about 500 ms. 2) He asked people to report when they became aware of an urge to move by noting the location of a dot moving around a clockface when they became aware of the urge (or intention or decision) to move. They typically reported awareness approximately 200 ms prior to movement. He reasoned that the RP was a sign of movement initiation, and argued that since it occurred several hundred milliseconds before awareness of the urge to move (let's call this the conscious intention to move), conscious intention could not be the cause of the initiation of movement. And if our conscious intentions do not cause our actions, then our actions are not freely willed.

Although Libet's methods have been widely criticized, the basic empirical results he reported have been replicated numerous times. An RP precedes spontaneously generated movement, and the average reported time of awareness, though prior to movement, nonetheless lags behind the beginning of the RP. Do the Libet results therefore demonstrate that the conscious will is inefficacious in generating action?

I think the answer is clearly *no*. The main problem is not with the empirical results, but with their interpretation. Here I mention only a few of what I consider to be the most salient criticisms of the interpretation of the Libet results. First, a careful look at Libet's methods reveals that the RP is generated from the average of many trials, time-locked to finger movement. Because one cannot discern the presence of an RP in a single trial because of the noisiness of the EEG data, one cannot determine whether every finger movement is preceded by a single-trial RP. But suppose it is. The bigger problem is that one cannot determine whether every single-trial RP is *followed by* a finger movement. That is, does the presence of the RP signify that an action is set in motion? Because Libet collected his data *only when finger movements occurred*, it is possible that single-trial RPs occur without motor action, sometimes or even often.[3] For instance, single-trial RPs may occur during anticipation of movement without movement, in cases when planned movements are aborted prior to initiation, and so on. If RPs occur without movement, then RPs are not an indication of action initiation, and so are not legitimately interpreted as a sign that the brain has initiated action independently of conscious will. Indeed, recent studies (Schegel et al., 2013) suggest that RPs can occur without being followed by movement, and some techniques are

[3] For another interpretation of the significance of Libet's data, see Schurger et al., 2012.

being developed that may allow us to identify individual RPs (Jose del R. Millan, personal communication).

Second, a number of studies have called into question the reliability of the clock-timing method for awareness. There is a high degree of variability in the reported time of awareness (Banks and Pockett, 2007) and a number of manipulations have shown that the reported time of awareness (the time of W) can be altered by factors that occur well after the action and after the subject must be aware of the stimulus, suggesting that W is at least in part determined retrospectively. This in itself may provide interesting fodder for philosophical theorizing, but with respect to the Libet studies, it strongly suggests that we maintain a pretty skeptical eye to the validity of the clock-timing method for measuring time of awareness. However, to my mind the strongest reason to discount Libet's interpretation is that I don't believe he is actually measuring time of conscious intention at all. To see why, let us consider Libet's experiment in closer detail.

The subject is asked to do two different tasks: (1) to spontaneously move his or her finger, and (2) to report where a rotating dot was on a clock face at the time he or she became aware of the intention to move it (W). (2) is a complicated task, and it has several components: (a) recognizing that one has an intention to move; (b) indexing the visual stimulus of the clock face when (a) is satisfied; (c) reporting the result of (b). It is likely that satisfying (a) requires attention to an internal state, and it is certain that satisfying (b) requires attention to an external stimulus, and it is quite likely that transitioning between (a) and (b) involves a shift of attention that takes time, causing the indexed visual stimulus of the clock time to be later than the actual time of occurrence of (a). That could explain some of the lag between RP and W. However, there is a more philosophical objection in the area, concerning the nature and content of the relevant states. Libet intends to measure the time of conscious intention to move. However, a closer look at Libet's experimental design suggests that this is not the state that he is actually measuring.

Notice, first, that in normal voluntary action over which we ordinarily think we have control (whether spontaneous action like a finger movement or actions that are in some way affected by, or responses to our environments, like reaching for a cup, stopping at a stop light, etc.), we are typically not aware of our *intention* to move or turn our hand or step on the brake such that we think, "Lo! Now I have an intention to [move, turn my hand, step on the brake]." If we did, our minds would be so cluttered with self-conscious thoughts that we would be hard-pressed to exert our agency in meaningful ways. Rather, we move, turn our hands, or step on the brakes intentionally, and could report having intended to do so if we were asked. Our intentions are conscious in that they are reportable, not in the sense that we are aware at the time of having them in the same way that we are with many perceptual stimuli. Our ordinary voluntary intentions are not phenomenologically present to us. Let us call these ordinary voluntary intentions (which I take to be paradigmatic cases of volition), C-intentions. The question is, does Libet measure the timing of C-intentions?

Answer: No. Libet is measuring a different state. In order to perform task (1) (spontaneously but deliberately moving a finger), the subject must form a C-intention that we may characterize as having the content "move finger" (or perhaps "move finger now"). In order to monitor one's conscious state and report on the timing of one's intention (task (2)), one

has to effortfully direct attention to one's C-intention, because intentions don't present themselves in the same way as perceptions. By so directing one's attention, we become aware that we have a C-intention, thereby forming a meta-state (W) with the content, "I am conscious of (having/being in) a state with the content 'move finger' ". It is this state that is indexed by the Libet-clock, a state different from, and moreover, dependent on, the C-intention. Because the state that is actually measured depends on the prior existence of the C-intention, and thus must occur after it, we can easily explain why W occurs relatively late with respect to the RP. This experiment, then, leaves open the temporal relation between the C- intention and the action.

What should we make of the Libet experiments? It is open to the philosopher to interpret C-intentions as conscious intentions, to maintain that the efficacy of conscious intentions is necessary for free will, but to deny that the Libet study measures the timing of conscious intentions. Or, it is open to the philosopher to accept that the W Libet measures is the timing of conscious intention (we might then call C-intentions "common intentions"), and to deny that free actions must stem from conscious intentions—rather, all that is necessary are common intentions. Either way, a better understanding of the experiments and their relation to philosophical issues shows that Libet's analysis is invalid as an analysis of the causal antecedents of free action; it is suitable only as an analysis of the timing of self-conscious action.[4]

A more recent study using fMRI has been conducted that purports to show something along the lines of the Libet study: that in a free decision task, a person's decision can be predicted seven to ten seconds before they decide, and thus before they are aware of their intention. Soon and colleagues (Soon, Brass, Heinze, and Haynes, 2008) asked people to decide whether to move a left or right finger while their brains were being scanned with fMRI. They reported that activity in a region in frontal cortex occurring seven to ten seconds before movement was predictive of which finger they would move. This was widely interpreted as showing that our brains decide before we do (Keim, 2008; Youngsteadt, 2008). However, a closer look at the data showed that the researcher's ability to predict on the basis of fMRI data was only slightly above chance: they could not "read off" the future action from the brain activity, nor did they have any indication that the future decision was already determined or could not be altered. Their work demonstrated that some prior information about the state of the brain reflected in fMRI activity was relevant to a future decision, but this is far from showing that our decisions are determined without "our" input. The finding is in fact not terribly surprising: the brain works at a multitude of timescales, and even memory of the last few decisions could have a bearing on our future (free) choices.

These examples of attempts to integrate neuroscience data with philosophical theorizing highlight the importance of having a firm grasp of both the neuroscience and the philosophy.

[4] Self-conscious action may also be important for discussions of freedom, for perhaps it is the case that some sorts of deliberations and weighings of reasons must occur self-consciously. Nevertheless, while our common-sense concept of freedom may require that our actions causally involve conscious intention, it is not clear that it requires self-conscious intention (or consciousness of conscious intention).

5.6 The Future: Areas in which Neuroscience May Influence Philosophy

To date, neuroscience evidence has played a role in debates about motive internalism, about deontology vs. consequentialism, about the role of emotion in cognition, about detecting awareness, and about criminal culpability. It has been implicated in numerous discussions about free will, and the role of awareness in free action. It seems likely that in the future, neuroscience evidence will increasingly be cited in work regarding control structures important for agency and responsibility, and in theorizing about consciousness. It will likely be used in establishing causal claims about mental phenomena, and will probably also play a role in future philosophical theorizing about innateness and learning, about mental representations, conceptual and nonconceptual content, and the fundamental building blocks of thought. The successful integration of neuroscience information into philosophical theorizing could improve both our philosophical and scientific understanding of the mind. But it will be essential both to understand the technical details of the neuroscience in order to assess what empirical conclusions can be drawn from them, and equally important to be clear about exactly what philosophically relevant object one takes these empirical results to speak to.

Acknowledgements

Thanks to Tamar Gendler and Ian Gold for comments on an earlier draft, to the John Templeton Foundation for support, and to the Mellon Foundation for summer funding, and to Hank Greely and the Stanford Law School for an excellent work environment.

References

Anderson, M. L. (2015). Mining the Brain for a New Taxonomy of the Mind. *Philosophy Compass* 10(1): 68–77.

Atkins v. Virginia, 536 U.S. 304 (2002).

Balaguer, M. (2004). A Coherent, Naturalistic, and Plausible Formulation of Libertarian Free Will. *Noûs*, *38*(3), 379–406.

Banks, W. P., and Pockett, S. (2007). Benjamin Libet's Work on the Neuroscience of Free Will. In M. Velmans and S. Schinder (eds.), *Blackwell Companion to Consciousness* (pp. 657–70). Malden, MA: Blackwell.

Berker, S. (2009). The Normative Insignificance of Neuroscience. *Philosophy & Public Affairs*, *37*(4), 293–329.

Brink, D. (1989). *Moral Realism and the Foundations of Ethics*. Cambridge: Cambridge University Press.

Cholbi, M. (2006). Belief Attribution and the Falsification of Motive Internalism. *Philosophical Psychology*, *19*(5), 607–16.

Craver, C. F. (2007). *Explaining the Brain*. New York: Oxford University Press.

Fodor, J. A. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, Mass: MIT Press.

Gazzaniga, M., Ivry, R., and Mangun, G. R. (2008). *Cognitive Neuroscience: The Biology of the Mind* (3rd Edition ed.). New York: W.W. Norton and Company.

Greene, J. D. (2008). The Secret Joke of Kant's Soul. In W. Sinnott-Armstrong (ed.), *The Neuroscience of Morality: Emotion, Brain Disorders, and Development* (Vol. 3, pp. 35–79). Cambridge, MA: MIT Press.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., and Cohen, J. D. (2004). The Neural Bases of Cognitive Conflict and Control in Moral Judgment. *Neuron*, *44*, 389–400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., and Cohen, J. D. (2001). An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science*, *293*, 2105–8.

Guillery, R. W. (2005). Observations of Synaptic Structures: Origins of the Neuron Doctrine and its Current Status. *Philos Trans R Soc Lond B*, *360*, 1281–1307.

Jones, O. D., Buckholtz, J., Schall, J. D., and Marois, R. (2009). Brain Imaging for Legal Thinkers: A Guide for the Perplexed. *Stanford Technology Law Review, 5*: 13–18.

Kahane, G. (2012). On the Wrong Track: Process and Content in Moral Psychology. *Mind & Language 27*(5): 519–45.

Kamm, F. M. (2009). Neuroscience and Moral Reasoning: A Note on Recent Research. *Philosophy & Public Affairs*, *37*(4), 330–45.

Kandel, E. R., Schwartz, J., Jessell, T., Siegelbaum, S., and Hudspeth, A. J. (2012). *Principles of Neural Science, Fifth Edition*. New York: McGraw-Hill.

Keim, B. (2008). Brain Scanners Can See Your Decisions Before You Make Them. *WIRED* (04.13.08).

Kennett, J., and Fine, C. (2007). Internalism and the Evidence from Psychopaths and "Acquired Sociopaths". In W. Sinnott-Armstrong (ed.), *Moral Psychology* (Vol. 3: *The Neuroscience of Morality*, pp. 173–90). Cambridge, MA: MIT Press.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., et al. (2007). Damage to the Prefrontal Cortex Increases Utilitarian Moral Judgments. *Nature*, *446*, 908–11.

Lehky, S. R., and Sejnowski, T. J. (1988). Network Model of Shape-From-Shading: Neural Function Arises from Both Receptive and Projective Fields. *Nature*, *333*, 452–4).

Millikan, R. (1984). *Language, Thought and Other Biological Categories: New Foundations for Realism*. Cambridge: MIT Press.

Owen, A. M., Coleman, M. R., Boly, M., Davis, M. H., Laureys, S., and Pickard, J. D. (2006). Detecting Awareness in the Vegetative State. *Science*, *313*(8 September), 1402.

Poldrack, R. A. (2006). Can Cognitive Processes be Inferred from Neuroimaging Data? *Trends in Cognitive Sciences*, *10*(2), 59–63.

Poldrack, R. A., and Wagner, A. D. (2004). What Can Neuroimaging Tell Us about the Mind? Insights from Prefrontal Cortex. *Current Directions in Psychological Science*, *13*(5), 177–81.

Pribram, K. H. (1982). Localization and Distribution of Function in the Brain. In J. Orbach (ed.), Neuropsychology after Lashley (pp. 273–96). New York: Erlbaum.

Roskies, A. L. (2003). Are Ethical Judgments Intrinsically Motivational? Lessons from "Acquired Sociopathy". *Philosophical Psychology*, *16*(1), 51–66.

Roskies, A. L. (2006). Patients with Ventromedial Frontal Damage have Moral Beliefs. *Philosophical Psychology*, *19*(5), 617–27.

Roskies, A. L., and Petersen, S. E. (2001). Visualizing Human Brain Function. In E. Bizzi, P. Calissano and V. Volterra (eds.), *Frontiers of Life: The Intelligent Systems* (Vol. III, pp. 87–109): Academic Press.

Schlegel, A., Alexander, P., Sinnott-Armstrong, W., Roskies, A. L., Tse, P. U. and Wheatley, T. (2013). Barking up the Wrong Free: Readiness Potentials Reflect Processes Independent of Conscious Will. *Experimental Brain Research 229*(3): 329–35.

Schurger, A., Sitt, J. D., and Dehaene, S. (2012). An Accumulator Model for Spontaneous Neural Activity Prior to Self-Initiated Movement. *Proceedings of the National Academy of Sciences 109*(42): E2904–13.

Smith, M. (1993). *The Moral Problem*. Oxford: Blackwell.

Soon, C. S., Brass, M., Heinze, H.-J., and Haynes, J.-D. (2008). Unconscious Determinants of Free Decisions in the Human Brain. *Nat Neurosci*, *11*(5), 543–5.

Treisman, A. M., and Kanwisher, N. G. (1998). Perceiving Visually Presented Objects: Recognition, Awareness, and Modularity. *Current Opinion in Neurobiology*, *8*(2), 218–26.

Ullman, M., Corkin, S., Pinker, S., Coppola, M., Locascio, J., and Growdon, J. H. (1993). Neural Modularity in Language: Evidence from Alzheimer's and Parkinson's Diseases. *Soc Neuro Abstr*, *19*, 1806.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., and Wager, T. D. (2011). Large-Scale Automated Synthesis of Human Functional Neuroimaging Data. [10.1038/nmeth.1635]. *Nature Methods*, *8*(8), 665–70.

Youngsteadt, E. (2008). Case Closed for Free Will? *Science NOW*. Retrieved from <http://news.sciencemag.org/sciencenow/2008/04/14-03.html/>. Accessed September 25, 2015.

CHAPTER 30

LOGIC AND PHILOSOPHICAL METHODOLOGY

JOHN P. BURGESS

1. Introduction

For present purposes "logic" will be understood to mean the subject whose development is described in Kneale and Kneale (1961) and of which a concise history is given in Scholz (1961). As the terminological discussion at the beginning of the latter reference makes clear, this subject has at different times been known by different names, "analytics" and "organon" and "dialectic", while inversely the name "logic" has at different times been applied much more broadly and loosely than it will be here. At certain times and in certain places—perhaps especially in Germany from the days of Kant through the days of Hegel—the label has come to be used so very broadly and loosely as to threaten to take in nearly the whole of metaphysics and epistemology. Logic in our sense has often been distinguished from "logic" in other, sometimes unmanageably broad and loose, senses by adding the adjectives "formal" or "deductive".

The scope of the art and science of logic, once one gets beyond elementary logic of the kind covered in introductory textbooks, is indicated by two other standard references, the *Handbooks* of mathematical and philosophical logic, Barwise (1977) and Gabbay and Guenthner (1983–9), though the latter includes also parts that are identified as applications of logic rather than logic proper. The term "philosophical logic" as currently used, for instance, in the *Journal of Philosophical Logic*, is a near-synonym for "nonclassical logic". There is an older use of the term as a near-synonym for "philosophy of language". This older usage is understandable, since so much of philosophy of language, and notably the distinction between sense and reference, did originally emerge as an adjunct to logical studies; but the older usage seems to be now obsolescent, and will be avoided here.

One side of the question of logic and philosophical methodology is that of the application of logic in philosophy. Since logic has traditionally been regarded as a methodological discipline, it is difficult or impossible to distinguish applications of logical *methods* from application of logical *results*, and no effort to maintain such a distinction will be made here. Distinctions and divisions within the topic of applications of logic

in philosophy are to be made, rather, on the basis of divisions of logic itself into various branches. Mathematical logic comprises four generally recognized branches: set theory, model theory, recursion theory, and proof theory, to which last constructive mathematics, not in itself really a part of logic but rather of mathematics, is attached as a kind of pendant. Philosophical logic in the relevant sense divides naturally into the study of *extensions* of classical logic, such as modal or temporal or deontic or conditional logics, and the study of *alternatives* to classical logic, such as intuitionistic or quantum or partial or paraconsistent logics: The nonclassical divides naturally into the *extra*classical and the *anti*classical, though the distinction is not in every case easy to draw unambiguously.

It should not be assumed that "philosophical logic" will inevitably be more philosophically relevant than "mathematical logic". Through the early modern period, logic as such was regarded as a branch of philosophy, but then that was equally the case for physics, and today the situation is quite different: Only a minority of professional logicians are housed in departments of philosophy, and this is true not just of specialists in "mathematical" logic but also of specialists in "philosophical" logic, many of whom are housed in departments either of mathematics or of computer science. Most nonclassical logics were initially introduced by philosophers, and with philosophical motives, but as their study has developed it has come to include the mathematical investigation of "logics" no one has ever advocated as accounts of the cannons governing deductive argumentation, just as geometry has come to include the mathematical study of "geometries" no one has ever seriously advocated as accounts of the structure of the physical space. For computer scientists, the literal truth of such philosophical ideas as may have played a role in motivating the original introduction of one or another logic is never what matters, but rather the heuristic suggestiveness and fruitfulness of such ideas, when taken in a perhaps metaphorical or unintended sense, for this or that technical application. The discussion to follow accordingly will not give special emphasis to philosophical logic merely because it is called "philosophical".

Rather, the seven branches of logic that have been distinguished—(1) elementary logic, (2) set theory, (3) model theory, (4) recursion theory, (5) proof theory, (6) extraclassical logics, (7) anticlassical logics—will be given roughly equal coverage. As it happens, each of the seven topic areas listed has a somewhat different flavor: The bearing of some branches on philosophy is pervasive, while the bearing of other branches is localized; the influence of some branches on philosophy has been positive, while the influence of other branches has been problematic; the relevance of some branches to philosophy is widely recognized, while the relevance of other branches is less known and imperfectly understood. As a result there is great variation in the nature of the philosophical issues that the involvement of the different branches with philosophy have raised. And as a result the discussion below will be something of a potpourri.

Philosophy of logic is as much to be distinguished from logic proper, including philosophical logic, as history of linguistics is to be distinguished from linguistics proper, including historical linguistics. Another side of the question of logic and philosophical methodology is therefore that of the methodology of philosophy of logic, insofar as it has a methodology of its own, distinct from the methodology of philosophy at large. The first question about special methods peculiar to philosophy of logic as distinguished from

other branches of philosophy is simply the question whether there are any such distinctive methods.

There is much to suggest that it ought to be answered in the negative. The scope and limits of philosophy of logic are quite differently understood by different philosophers of logic, as comparison of such classics as Strawson (1952) and Quine (1970), not to mention Haack (1978), soon reveals. But a not-too-controversial list of central topics in present-day philosophy of logic might include the following: *Should truth-bearers be taken to be sentence types, or sentence tokens, or propositions; and if the last, are these propositions structureless or structured; and if structured, are they coarse-grained and "Russellian" or fine-grained and "Fregean"? Are logical forms the same as grammatical forms, or perhaps the same as "deep" in contrast to "surface" grammatical forms; and whether or not they are, are they psychologically real, represented somehow in the mind or brain of the reasoner, or are they merely imposed by the analyst in the course of evaluating reasoning? Does the source of logical truth and logical knowledge lie in the meanings of the logical particles or elsewhere; and should that meaning be conceived of as constituted by truth conditions or by rules of use?* Obviously these central questions of philosophy of logic are very closely linked to central questions of philosophy of language and/or philosophy of linguistics. Indeed, they are so closely linked as to make it hard to imagine how there could be methods peculiar to philosophy of logic alone and not relevant also to these or other adjoining fields.

Yet upon further reflection it appears that there is after all at least one special methodological puzzle in philosophy of logic that may be without parallel elsewhere. The problem in question arises in connection with philosophical debates between propopents of anticlassical logics and defenders of classical logic, and it amounts to just this: What logic should be used in evaluating the arguments advanced by adherents of rival logics as to which logic is the right one? It is natural to suspect that both sides would soon become involved in circular reasoning; no doubt one side would be arguing in a vicious circle and the other side in a virtuous one, but still the reasoning would be circular on both sides. The question of how if at all noncircular debate over which is the right logic might be possible is perhaps the most readily identified distinctive methodological problem peculiar to philosophy of logic. It can, however, conveniently be subsumed under the question of the role of anticlassic logics in philosophy, which is already on the list of seven topics for exploration we have enumerated.

2. Elementary Logic and Philosophy

Elementary logic, of which the half-dozen branches of advanced mathematical and philosophical logic that have been identified are so many specialized outgrowths, is concerned with the evaluation of arguments, but not just any kind of argument and not just any kind of evaluation. Its concern is with deductive arguments, arguments purporting to show that, assuming some things, something else then follows conclusively and not just probably. And its concern is with the formal validity of such arguments, with whether the forms of the premises and conclusion guarantee that if the former are true the latter is so as well, and not with their material soundness, with whether the premises are as a matter

of actual fact true. Now in philosophy today the greatest variety of methods are employed. Nonetheless, deductive argumentation remains what it always has been, a very important and arguably the single most important philosophical method. Though it is impossible to collect precise statistics on such questions, undoubtedly philosophy remains among intellectual disciplines the second-heaviest user of deductive argumentation, next after mathematics but ahead of jurisprudence, theology, or anything else. And though formal validity is only one virtue to be demanded of deductive argumentation, it is a very fundamental and arguably the single most fundamental virtue, the *sine qua non*. Accordingly, it is widely agreed that every student of philosophy needs a least a rudimentary knowledge of logic, of how to assess the formal validity of deductive arguments. The point is perhaps not *universally* agreed: It would presumably be disputed by Andrea Nye, since Nye (1990) reaches the conclusion that "logic in its final perfection is insane"; but this is a radical—one may even say fringe—position.

What is more often disputed is not that students of philosophy should have a modicum of practical knowledge of logic, but rather how much is enough. How many concepts, how much terminology, must the student take in? Certainly the student needs to possess the concept of an *argument* in something like Monty Python's sense of "a connected series of statements intended to establish a proposition" as opposed to the colloquial sense of "a loud, angry exchange of opinions and insults". Surely the student also needs to understand the distinction between formal *validity* and material *soundness*—and it should be added, needs to appreciate the chief method for establishing *in*validity, that of exhibiting a parody, another argument of the same form whose premises are manifestly true and whose conclusion is manifestly false. (This is the method illustrated by the Mad Hatter when he replies to the assertion that "I mean what I say" and "I say what I mean" are the same, by objecting that one might as well say that "I see what I eat" and "I eat what I see" are the same. It is also the method used by Gaunilo replying to Anselm.) Ideally, the student should know some of the labels used in describing the logical forms of premises and conclusions, and for some of the most common kinds of valid arguments, and for some of the most egregious fallacies: Terms like *biconditional* and *modus ponens* and *many questions* should be in the student's vocabulary. (At the very least, the student should know enough to avoid the illiterate misuse of the expression "beg the question" that has become so annoyingly common of late.) But how much more should the student know? And is there any need to initiate the student into the mysteries of logical symbolism?

There is then also a further question about *how* the student should acquire the range of knowledge called for, whatever its extent may be. Undergraduate concentrators in mathematics, who all at some fairly early stage in their training need to "learn what a proof is", generally do so not through the explicit study of logic, but in connection with a course on some core branch of mathematics, perhaps on number theory, perhaps real analysis (calculus done rigorously); if they undertake a formal study of mathematical logic, as most do not, it will be at some later stage. Perhaps, then, the modicum of logical vocabulary and theory needed by students of philosophy should likewise be imparted, not in a separate course, but in conjunction with some kind of introductory topics-in-philosophy course. Or perhaps it should be left to writing courses, except that one hears horror stories about what students are told in such courses ("Your writing is much too clear") by instructors from

literature departments who are under the baleful influence of certain fashionable theoreticians. In short, while surely some course in elementary, introductory-level logic should be offered, what is debatable is whether it should, for prospective philosophy concentrators, be made a requirement or left as an elective.

There is then also a further question about what the content of such a course, whether required or not, should optimally be, and in particular, what additional material should be included beyond the modicum of formal, deductive logic that is absolutely essential. Should it just be more formal, deductive logic? Or should it be a bit of what is called "informal logic", or critical thinking? Or should it be a bit of what is called "inductive logic", or probabilistic reasoning? Or should it be "deviant logic", or anticlassical positions? Or should it be a little of this and a little of that? The appearance of this volume suggests still yet another alternative, that of folding instruction in elementary logic into a general "methods of philosophy" course. The main point is that the most obvious issues raised by the role of elementary logic in philosophy are curricular issues, affecting the philosopher qua teacher of philosophy more than the philosopher qua philosopher.

3. Set Theory and Philosophy

Sophisticated developments in higher axiomatic set theory (as described in Part B of Barwise [1977]) have influenced philosophy of mathematics, but treatment of the matter will be postponed so that it may be discussed in conjunction with the influence of proof theory on that same specialized branch of philosophy. Leaving all that aside for the moment, more elementary set theoretic results—or if not results, at least notation and terminology—are quite commonly used in a variety of branches of philosophy, as they are quite commonly used in a variety of branches of many other disciplines. The most elementary set-theoretic material, including such concepts as those of *element, subset, intersection, union, complement, singleton, unordered pair*, and *ordered pair*, the material whose use is the most widespread in philosophy, has penetrated instruction in mathematics down to the primary-school level, and can be presumed to be familiar to students of philosophy without much need for separate discussion, except perhaps a very brief one to fix notation, which has not been absolutely standardized. In many branches of analytic philosophy, however, a bit more of set theory is involved. One may go on to use some marginally more advanced notions, perhaps those pertaining to certain special kinds of binary relations such as *functions* and *orders* and *equivalences*, along with attendant concepts like those of *injectivity* and *surjectivity* and *bijectivity*, or of *reflexivity* and *symmetry* and *transitivity*. There may also be some need or use for the notion of *ancestral* from what is called "second-order logic", a part of the theory of sets or classes that sometimes passes for a branch of logic. And not all of these matters can be counted on to have been already absorbed by students in primary or secondary school mathematics. But the problems raised by the role of set theory in philosophy are not exclusively the kinds of curricular issues that we have seen to arise in connection with the role of elementary logic (though indeed *if* the introductory logic curriculum is to be rethought, the possible

introduction of a bit more set theory than is customarily covered at present might be one issue to be considered).

One significant problem raised by philosophers' use of set-theoretic notions and notations is an embarrassment that arises for philosophers of a certain bent, those inclined towards "nominalism" in the modern sense. For views of this kind have no patience with and leave no room for sorts of entities for which it makes questionable sense to ask after their location in time and space, and no sense to ask after what they are doing or what is being done to them. And sets are paradigmatic examples of entities that are of such a sort, often pejoratively called "Platonic", historically absurd though this usage is, or more neutrally called "abstract". Philosophers inclined to nominalistical views will, it seems, need to watch out and take care that they do not, in the very exposition and development of those views, fall into violations of their professed principles by making mention, in the way that is so common among philosophers, of abstract, so-called Platonic apparatus from set theory. For opponents of nominalism have often argued that if would-be nominalists can be caught frequently using set-theoretic notions themselves, then such notions cannot really be so intellectually disreputable as nominalist doctrine would maintain, and acquiring knowledge of them cannot really be so impossible as popular epistemological arguments for nominalism insist. The conflict between the widespread use of set theory within and outside logic, and nominalist challenges to abstract ontology is taken to be *the* main problem in philosophy of logic in Putnam (1971), the *locus classicus* for the "indispensability argument", according to which, set theory being useful and used in logic, mathematics, science, and philosophy to the point that one could hardly do without it, one ought simply to accept it. But the issues seem today by no means so clear-cut as they did to Putnam.

For proponents of nominalism today often imagine there is some cheap and easy solution to the difficulties of the philosopher who would like to pose as a hard-headed nominalist without having to give up the use of any of the customary set-theoretic and mathematical methods that have penetrated into contemporary analytic philosophy. The supposed solution is to be found in some kind of instrumentalism that will allow them to use set-theoretic language when speaking out of one side of their mouths, while continuing to deny the existence of sets when speaking out of the other side; or else in some kind of distinction that will allow them to say sincerely that it is literally true that sets exist, while still denying that they have thereby undertaken any "ontological commitment" to sets. Sympathizers with nominalism now point to new "fictionalist" possibilities, permitting one to use in practice whatever is useful while still rejecting it in principle—while critics complain that waters previously clear have been muddied by obscurantism about a supposed gap between "existential implications" and "ontological commitments". But these contentious issues are all too familiar to those who follow the literature in metaphysics and philosophy of mathematics, and need not be enlarged upon further here.

4. Model Theory and Philosophy

Alfred Tarski's work on model theory (the foundation stone of the subject as expounded in Part A of Barwise [1977]) arose out of his famous definition of truth. The strategy used in Tarski (1956) was the method of *giving a characterization of what it is for a sentence of a language to be true under a given interpretation by induction on the syntactic complexity*

of the sentence, for instance, defining truth for a conjunction in terms of truth for its conjuncts. The same method is adopted and adapted also in Kripke (1963) to give a model theory for modal and related logics, which involves introducing a system of indices, picturesquely called "possible worlds", and the relativizing of the notion of truth to that of "truth at a possible world". Both the method of definition by induction on complexity and the notion of possible world have become immensely influential, the former especially in the philosophy of language of Donald Davidson, the latter especially in the metaphysics of David Lewis (see Lewis [1986]). Yet the closer one looks at the original work of Tarski and Kripke, the more dubious becomes the supposed connection between that work and the developments in philosophy of language and metaphysics that it has somehow given rise to. In the case of the metaphysics of possible worlds, the looseness of the connection is generally recognized, since Kripke notoriously very explicitly and emphatically repudiated anything like the Ludovician conception of possible worlds as something like distant planets way off in logical space. In the case of philosophy of langauge, the looseness of the connection between "formal semantics" or model theory and "linguistic semantics" or meaning theory is perhaps not so widely understood.

It seems to be not so widely recognized as it might be that truth-conditional theories of meaning as developed by Davidson and others represent an inversion rather than an application of the Tarskian standpoint: Tarski took *truth* to be the problematic notion, rendered suspect by the well-known paradoxes, whose meaning needed explanation or definition, and took as understood and available for use in his definition the meanings of the expressions of the language for which truth was being defined; whereas Davidson takes the notion of *truth* more or less for granted as an unanalyzed and undefined primitive, and attempts to use it to characterize the meanings of expressions of the object language. Davidson himself was quite self-consciously turning Tarski on his head, but Davidson's followers have perhaps not always recognized that truth-conditional semantics is not Tarski rightside up but Tarski upside down.

Tarski did call model theory "semantics", as indeed did Kripke; but what Tarski meant by "semantics" is not at all what linguists and philosophers of language today mean by it, as should be clear enough, even without going into the complicated history of the usage of the term, from the fact that Tarski's list of paradigmatically "semantic" notions includes *truth* but not *synonymy*. The thought that truth-conditional semantics of a Davidsonian kind (or with variations of a Kaplanian kind) is anything like a direct application of "formal semantics" of a Tarskian kind (or with variations of a Kripkean kind) is simply mistaken, and represents a kind of fallacy of equivocation on the ambiguous term "semantics". Whether the departure from or inversion of the model-theoretic perspective that has led to truth-conditional theories of meaning was a good thing or a bad thing is too large an issue to be entered into here; but departure or inversion it unquestionably was: Logic can neither take the credit nor bear the blame for truth-conditional semantics.

5. Recursion Theory and Philosophy

The consortium of disciplines collectively known as "cognitive studies" includes among other components philosophy of mind, neurology, and several branches of computer science. The whole subject of computer science is, on its theoretical side, an outgrowth of the branch of logic called "recursion theory" (as expounded in Part C of Barwise [1977]), now

sometimes alternatively called "computability theory", along with its offshoot, complexity theory. As a result, acquaintance with the basics of this branch of logic—with the notion of *Turing machine*, above all—is desirable if not indispensable background for philosophers involved in cognitive studies. One needs this kind of background simply to read a lot of the current literature, both the large positive literature that endorses and in various ways applies a "computational theory of mind" and the much smaller negative literature that argues there are deep conceptual confusions in "machine-state functionalism".

The positive literature is too vast to be intelligently surveyed in the space available here; nor is the present author the best person to undertake such a survey. The relation of computer science to philosophical methodology really deserves a chapter of its own, or perhaps even two chapters (with the branch known as "artificial intelligence" getting separate treatment). Central to the much smaller negative literature is the discussion of "Kripkenstein's skeptical paradox" (Kripke [1982]), and even that is too large a subject to be gone into seriously here. The fundamental problem is just this: What is it for some physical object to constitute a realization of some abstract algorithm, or for some material organ such as a brain to be an embodiment of an idealized machine? The same object may be construed as an imperfect realization of any number of different abstract algorithms, and there is nothing *in the object* to make one construal correct and the others erroneous. Or so Kripke and followers argue. As Kripke has noted, this problem, if it is a genuine one, creates difficulties, not only for functionalist philosophy of mind and functionalist cognitive psychology, but also for much of contemporary philosophy of language and contemporary linguistics, which following Chomsky makes use of a notion of "competence" that is not to be identified with observable "performance" but is nonetheless supposed to be "psychologically real". The background in the pertinent branch of logic, recursion theory, that one needs to follow the discussion in either the positive or the negative literature is not extensive, and perhaps could be obtained from popularizations, without the need to enter deeply into technicalities; but background there is.

Recursion theory's centerpiece, the Church–Turing thesis, is relevant to philosophical methodology in quite another way. The thesis is important not only as an analytical tool, but also as a paradigm of the successful solution to a difficult problem of analysis. The problem Alonzo Church and Alan Turing addressed was the following. Impossibility results in mathematics have a certain utility in telling us not to waste time attempting certain tasks, though how much value such a warning has will depend on what use is made of the time saved. Negative, impossibility, results almost always require more background analysis than positive, possibility, results. If one wants to show it is possible to construct a given figure with ruler and compass, it is enough to give instructions for the construction, and a proof that it works as advertised. If one wants to show a construction *im*possible, however, one needs some sort of analysis of what constructibility amounts to. As with constructibility, so with computability. If one wants to show a function is computable, it is enough to present the instructions for computing it, and a proof that they work. If one wants to show a function is *un*computable, however, one needs some sort of analysis of what computability amounts to. Church and Turing each undertook, independently of the other, this task of analysis, seeking to find a rigorously definable mathematical notion that would be coextensive with the intuitive notion of computability. Church proposed to identify computability with (something called lambda-calculability, and later with) recursiveness, and Turing

with computability by one of his ideal machines. It was quickly seen the functions computable by a Turing machine are precisely the recursive functions, so that the theses of Church and Turing are in a sense equivalent. They are now almost universally accepted by experts. There is, however, a certain difference.

There is no hope of giving a fully, formally rigorous mathematical proof of the coincidence between some rigorously defined notion and some intuitive notion, since all the notions involved in a fully, formally rigorous mathematical proof must be rigorously-defined and not intuitive ones. (That is one reason why there are so few fully, formally rigorous proofs in philosophy as compared to mathematics.) So neither Church nor Turing offered a fully, formally rigorous mathematical proof for his thesis, nor did either claim that "computable" just *means* recursive or Turing computable, or that his thesis was analytic. Turing, however, did offer a heuristic argument, based on the thought that all one ever does in a computation is make and erase marks and move around the page, and that making and erasing marks could be done one stroke at a time, and moving around the page one step at a time. Church, by contrast, did little more than cite the nonexistence of any obvious counterexamples (and in the seven or eight decades since his day, no one has since found a plausible one). Church's thesis and Turing's thesis thus represent apparently extensionally successful analyses that can hardly be claimed to be analytic, and Church's work and Turing's work exhibit two different ways, one more a posteriori and one more a priori, in which one could hope to argue for such an analysis. There is a lesson for analytic philosophers in all this, about the scope and limits of the method of analysis, but it is perhaps one that, requiring as it does familiarity with some rather technical material, has not as yet been as widely understood or as seriously taken to heart as it should be.

6. Proof Theory and Philosophy

Proof theory in the narrow and strict sense (in which it is understood in Part D of Barwise [1977]) consists of certain specific types of theorems about certain specific types of formalisms: cut-elimination theorems for sequent calculi in the style of Gerhard Gentzen, and normalization theorems for systems of natural deduction in the style of Dag Prawitz. Its techniques and its technicalities have to a degree been brought by Michael Dummett and his followers into debates about classical *vs.* intuitionistic logic, but it is not that side of proof theory that I wish to consider here. I will be concerned rather with proof theory in a broader and looser sense, the kind of study that begins with Kurt Gödel's two famous incompleteness theorems, which are the main goals of any intermediate-level course in logic, and are treated in many textbooks (often in conjunction with the Church–Turing thesis), besides being the subject of a large literature of popularization of very mixed quality. The second of the two theorems says, roughly speaking, that if one restricts oneself to the most constructive and least controversial means of proof, then one cannot prove any *absolute* consistency results for any interesting mathematical axiom systems: one cannot prove that any such system is free from contradiction. One can, however, often prove *relative* consistency results, to the effect that system B is consistent relative to system A, meaning that if system A is free from contradiction, then so is system B, or contrapositively, if

there is a contradiction in system B, then there is a contradiction in system A. Proof theory in the broader or looser sense is concerned with comparing the "consistency strengths", where B counts as being of the same consistency strength as A if B can be proved consistent relative to A and vice versa, while B counts as being of lesser consistency strength than A if B can be proved consistent relative to A, but not vice versa.

Logicians have shown that virtually all systems that have been seriously proposed as foundations for mathematics fall somewhere on a linearly ordered scale of consistency strengths that leads from a very weak but still nontrivial system called "Robinson arithmetic", to a very strong system called "Zermelo-Frankel set theory plus rank-into-rank large cardinals". This itself is a striking result, since it is very easy to contrive artificial examples of systems that are of incomparable consistency strength. Why then should there be no naturally occurring examples? This basic result, the work of many hands, is accompanied by any number of other striking theorems. Some of these results, due to various workers, show that the bulk of the mathematics that finds serious applications in science and engineering can be developed in systems quite low down in the scale, where dwell most of the systems proposed by dissident "constructivist" mathematicians. Another one of these results, due to Yuri Matiyasevich building on work of several predecessors, shows that every time one moves up a notch on the scale, more theorems of number theory of a very simple type (asserting the non-existence of solutions to a certain Diophantine equation) become provable. The most sophisticated methods of proof theory in the narrow and strict sense are used in the study of the lower end of this scale, while sophisticated methods of a quite different, set-theoretic kind (notably Paul Cohen's method of forcing) are used at the upper end.

Many would say the results obtained by such methods are of considerable potential relevance to issues about mathematics that are much debated among philosophers. They would add that it is unfortunate that knowledge of such results, which in itself does not require deep involvement in the technicalities of their proofs, is perhaps not as widespread as it ought to be (to the extent that philosophers can sometimes be found writing as if they believed, contrary to the Matiyasevich theorem, that pie-in-the-sky set-theoretic assumptions about "large cardinals" just *cannot* have any impact on anything so down-to-earth as number theory). Unsurprisingly, given that in philosophy everything is potentially disputable and almost everything is actually disputed, there are others who would discourage study of the results from mathematical logic I have been discussing, or for that matter any other results from mathematical logic at all. They will perhaps quote *dicta* of Wittgenstein about the "disastrous invasion" of mathematics by logic, and about "the so-called mathematical foundations of mathematics" being merely a painted rock under a painted tower. As in other cases, I can here only note the existence of a disagreement over the role of logic in philosophy, without attempting to resolve it.

7. Extraclassical Logics and Philosophy

The traditional logic, based on Aristotle's syllogistic, was inadequate to the task of analyzing serious mathematical proofs, mainly because it lacked any treatment of relations. The logic that has displaced traditional syllogistic and that is now called "classical" was

developed, mainly by Gottlob Frege, precisely for the purpose of analyzing mathematical reasoning. Classical logic goes beyond traditional logic by just as much as is needed to analyze mathematical arguments—just as much, and no more. It takes no note of grammatical mood or tense, of epistemic or deontic modalities, or of subjunctive or counterfactual conditionals, since none of these matter for mathematics or are to be found in purely mathematical language. By contrast, they are very much to be found in philosophical language, and do matter for philosophy. Hence there would seem to be much room for philosophically relevant extensions of classical logic (as treated in Gabbay and Gunthner [1984]), enriching its formal language with the sorts of things just enumerated.

Modal logic in the broad sense—comprising temporal logic, epistemic and deontic logic, conditional logic, and more, as well as modal logic in the narrow sense of the logic of "necessarily" and "possibly", which is the part of the larger subject that will be of most concern here—has aspired to provide just such philosophically relevant extensions of classical logic. It was largely developed in hopes of making itself philosophically useful. Performance, however, has not lived up to promise, and it is not going too far to say that at times modern modal logic has done more to darken rather than to enlighten our understanding. Some of the most glaring deficiencies of conventional modal logic were early pointed out by the hostile critic Quine, but unfortunately the reaction of modal logic's champions to Quine's critique was highly defensive and often uncomprehending. More was done in the way of developing elaborate technical constructions to prove various conventional systems to be formally consistent, than was done in the way of analyzing and explaining the notions of necessity and possibility and their representation in language in order to show the systems in question intuitively intelligible, or where they were not so, to replace them by novel systems that would be.

One reason modal logic has been little able to provide guidance to philosophers engaged in modal reasoning is that modal logicians have never been able to agree as to which modal logic is the right one. And no wonder, since prior to Kripke (1972) they were generally hopelessly confused about the nature of necessity and possibility: The possible in the sense of what potentially could have been was conflated with the possible in the sense of what can without self-contradiction be said actually to be. As we now say, "metaphysical" modality was conflated with "logical" modality. But even today, when the importance of that distinction has been widely, though by no means universally recognized, modal logic is still full of *dubia*, even at the sentential level. The state of quantified modal logic (QML) is much worse. Until Kripke (1963), modal logicians did not even know how to develop systems of QML that would avoid—in the sense of making them optional extras, that one can assume or not as one chooses, rather than something built in to the basic formalism—the dubious "Barcan formulas", which imply that anything that could possibly have existed actually does exist, and that nothing that actually does exist could possibly have failed to exist. Even with that problem out of the way, however, the basic syntax of conventional QML is out of alignment with the way in which modal distinctions are expressed in natural language. For the formalism treats a modality as operating on a whole clause, including all of its subordinate clauses, while in natural language, modal distinctions operate on verbs, allowing the grammatical mood of a subordinate clause to differ from that of the main clause to which it is subordinate. The result is that with the conventional formalism it is difficult or impossible to express something like "If all those who wouldn't have come here if they hadn't been

obliged to do so now leave, there will be no one left" or "I could have been a lot thinner than I am"; nor is the addition of "actuality operators" to the language a sufficient remedy. There are a number of philosophical logicians at work today trying to improve matters, but it is too early to say whether a genuinely philosophically useful modal logic is going to emerge from their efforts.

To look briefly on the bright side, workers in theoretical computer science do seem to have found a number of modal systems useful in *non*-philosophical ways. Also, the category of extraclassical logics is not exhausted by modal logic, even when "modal logic" is taken in the broadest sense, and a number of extraclassical but nonmodal logics, including plural logic and predicate-functor logic, have occasionally figured in interesting ways in philosophical projects, though there is no space to go into such matters in detail here.

8. Anticlassical Logics and Philosophy

A. W. Kinglake is said to have proposed that every church should bear over its doors the inscription "important if true". Whatever one thinks of that suggestion, there is no question but that the three words of the proposed inscription apply to the claims made by proponents of anticlassical logics (as surveyed in Gabbary and Guenthner [1985] as well as Haack [1978]). The advocates of paraconsistent logics may go furthest in making claims about how many philosophical problems would be easily solved if their principles were adopted, but advocates of other anticlassical logics are not far behind. Further, for a number of such logics, some applications of their technical formalisms have been suggested that would *not* require literal belief in the underlying motivating philosophical ideas, giving another potential reason to study the formalisms, even if one rejects the ideas. Moreover, a large number of technical results, some of them quite impressive as pure mathematics, have accumulated concerning such logics. A curious phenomenon, however, may be observed in the technical literature on the metatheory of anticlassical logics.

With the exception of the mathematical intuitionists, advocates of anticlassical logics generally *make no serious effort to conform their own metatheoretic reasoning to patterns that are valid according to their own professed views*. Their own deductive behavior thus suggests that they secretly believe in the classical logic they profess to reject. This phenomenon was perhaps first noted by Kripke (unpublished) in the work of "relevance" or "relevant" logicians, who officially declare the inference "P or Q, but not P, so Q" to be "a simple inferential mistake such as only a dog would make," but who nonetheless were caught by Kripke using that very forbidden form of inference in the proof of a major metatheorem. How far it is legitimate for defenders of classical logic to invoke this curious phenomenon in debating with attackers is itself a debatable question. On the one hand, many writers on informal logic or critical thinking would hold it to be a fallacy to argue that since the heretics can't themselves live up to their doctrines, those doctrines must be in error; and it is indeed not entirely inconceivable that human thought should be drawn irresistably into certain patterns of inference that are nonetheless hopelessly wrong. In a way and in a sense, some of the psychological literature on heuristics points in something like such a direction. Nonetheless, it is less common for advocates of anticlassical logics to respond in this way

to Kripke-style objections than for them to respond that even though classical logic is not to be relied upon in general, there are special reasons why it may be relied upon in certain special areas, including the metatheory of anticlassical logic. Extended but inconclusive debates about "classical recapture" then ensue.

On the other hand, given the great difficulty, alluded to in the introductory remarks at the outset above, of non-question-begging debate over logical principles, it is very natural to slide from the question "Which logic is right?" to the question "Which logic should we follow?" And in connection with the latter question the observation that the logic one's opponents are proposing we should follow is one that they are incapable of following themselves is a perfectly cogent objection. After all, "ought" implies "can", does it not? In practice, debate often slides further from the prescriptive question "Which logic *should* we follow?" to the descriptive question "Which logic *do* we follow?" as both sides appeal to common sense. Be all that as it may, we still have the curious phenomenon noted before us.

With the exception of the mathematical intuitionists, most advocates of anticlassical logics are vulnerable to an objection, or anyhow subject to an observation, that Solomon Feferman has emphasized in connection with partial logic (a three-valued logic with "truth value gaps"): "Nothing like sustained ordinary reasoning can be carried out" using the logic they propose. The reason why this should be so is not immediately clear, but that it is so is what experience in working with these logics, along with the evidence of anticlassical logicians' metatheoretic behavior, suggests. Even the mathematical intuitionists may be no real exception. The intuitionists have been able to develop a substantial intuitionist or constructivist mathematics entirely in conformity with their principles, with all proofs conforming to intuitionistic logic; and of this intuitionistic or constructivistic mathematics, the intuitionists' work on the metatheory of intuitionistic logic forms one chapter. But the intuitionists have made no serious, sustained attempt to develop an intuitionist empirical science, and it is difficult to make out what logic, precisely, their underlying principles would imply to be the appropriate one for empirical reasoning.

For intuitionists or neo-intuitionists, in explaining why they adhere to the logic they do in mathematics, cite issues about proof: They take the very meaning of mathematical statements to be constituted by their proof-conditions. But in the empirical domain there is generally no question of apodictic proof. Evidence may establish a presumption, but presumptions are always defeasible, giving empirical reasoning a nonmonotonic character: What one is warranted in asserting given certain evidence may become unwarranted given more evidence. Moreover, performing the operations that would be needed to verify or falsify one empirical claim may make it impossible to perform the operations needed to verify or falsify another, and this phenomenon is by no means confined to the microscopic world or to quantum interference. These features, which have no counterparts in mathematics, suggest that a different logic from the one intuitionists use in mathematics would be required; but the details of the required logic have not been worked out.

Even if one is not at all tempted to adopt an anticlassical logic, the mere thought that someone might do so, or that some planet may harbour intelligent extraterrestrials who have done so, and whose mathematics and science must therefore be very different from ours, is philosophically intriguing, suggesting as it does a very radical form of "underdetermination of theory by evidence" or "conventionality". This is one way in which anticlassical

logics can exercise a fascination even over philosophers who are by no means willing to give up classical logic.

9. Conclusion

The foregoing hodgepodge of affirmations, denials, and interrogatives cannot be summed up in any simple slogan. From the miscellany of observations, reflections, and evaluations offered above, no single clear message, no overarching moral about logic and philosophy readily emerges. Examination of the many areas where the large and diversified field of logic impinges on the even larger, even more diversified field of philosophy merely confirms what was said at the outset, that the role of some parts of logic in philosophy is pervasive and positive, while that of others is peripheral, and that of yet others problematic. But at least one may say this: The interaction of logic with philosophy remains after more than two millenia a lively ongoing process. And one may also add this: If logic is to have the fullest possible positive influence, there will be a need both for existing achievements of logic to be more effectively communicated to philosophers and for logicians to expand and extend, and in some cases alter and amend, their own work.

References

Barwise, K. Jon, ed. (1977) *Handbook of Mathematical Logic*, Amsterdam: North Holland.

Gabbay, Dov and Franz Guenthner, eds. (1983) *Handbook of Philosophical Logic*, vol. 1, *Elements of Classical Logic*, Dordrecht: Reidel.

Gabbay, Dov and Franz Guenthner (1984) *Handbook of Philosophical Logic*, vol. 2, *Extensions of Classical Logic*, Dordrecht: Reidel.

Gabbay, Dov and Franz Guenthner (1985) *Handbook of Philosophical Logic*, vol. 3, *Alternatives to Classical Logic*, Dordrecht: Reidel.

Gabbay, Dov and Franz Guenthner (1989) *Handbook of Philosophical Logic*, vol. 4, *Topics in the Philosophy of Language*, Berlin: Springer.

Haack, Susan (1978) *Philosophy of Logics*, Cambridge: Cambridge University Press.

Kneale, William and Martha Kneale (1962) *The Development of Logic*, Oxford: Oxford University Press.

Kripke, Saul (1963) "Semantical Considerations on Modal Logic", *Acta Philosophica Fennica* 16: 83–94.

Kripke, Saul (1972) "Naming and Necessity: Lectures Given to the Princeton University Philosophy Colloquium", in Gilbert Harman and Donald Davidson (eds.), *Semantics of Natural Language*, Dordrecht: Reidel, 253–355 and (Addenda) 763–9.

Kripke, Saul (1982) *Wittgenstein on Rules and Private Language: An Elementary Exposition,* Cambridge: Harvard University Press.

Lewis, David Kellogg (1986) *On the Plurality of Worlds*, Oxford: Blackwell.

Nye, Andrea (1990) *Words of Power: A Feminist Reading of the History of Logic*, London: Routledge.

Putnam, Hilary (1971) *Philosophy of Logic*, New York: Harper & Row.

Quine, Willard Van Orman (1970) *Philosophy of Logic*, Englewood: Prentice-Hall.
Scholz, Heinrich (1961) *Concise History of Logic*, New York: Philosophical Library.
Strawson, Peter F. (1952) *Introduction to Logical Theory*, London: Methuen.
Tarski, Alfred (1956) "The Concept of Truth in Formalized Languages", trans. J. H. Woodger, in Tarski, *Logic, Semantics, Metamathematics: Papers from 1923 to 1938*, Oxford: Clarendon Press, 152–278.

CHAPTER 31

PHILOSOPHY OF MATHEMATICS

Issues and Methods

STEWART SHAPIRO

1. Introduction

I doubt that philosophy of mathematics is an intellectual Natural Kind, with sharply defined boundaries, and so I do not feel wedded to substantial theses concerning what does and what does not count as pursuing this discipline. In this survey, I sketch a number of issues and problems that motivate at least much of the literature in the philosophy of mathematics, but without any claims to be comprehensive.[1] This sets the stage for various matters of methodology.

2. Metaphysics, Epistemology, Semantics

The primary purpose of the philosophy of mathematics is to *interpret* mathematics, and illuminate the place of mathematics in the overall intellectual enterprise. There are, first, ontological questions. What is the subject matter of mathematics, if it has one? In particular, what *are* numbers, points, sets, functions, arrows, nodes, and the like, if such things exist (and if they are "things")? Of course, mathematics itself has something to say about its objects: sets have elements; functions have domains, codomains, arguments, and values; numbers have successors; sets have powersets; etc. The philosopher is more concerned with the metaphysical nature of these mathematical objects (again, if there are any) and how they compare to other sorts of objects.

[1] For a fuller account see, for example, Shapiro (2000, chapters 1–2). Even there, I do not claim to have provided complete coverage.

These ontological questions give rise to a number of positions. *Realism in ontology* is the view that at least some mathematical objects exist objectively, independent of the mathematician (as held by, e.g. Gödel [1964], Maddy [1990], Shapiro [1997]). So numbers, for example, are the same sort of things as trees and baseballs. Realism in ontology is sometimes called "Platonism" or "platonism", after a similar view, held by Plato, concerning Universals and the like (although Plato seemed to hold that the abstract Forms are *more* real than ordinary, physical objects). The main alternatives to ontological realism are varieties of idealism and nominalism. The idealist agrees with the realist that mathematical objects exist, but adds that they depend on the (human) mind. They do not exist objectively. One theme is that mathematical objects are constructs arising out of mental activity. It seems clear, however, that mathematics is common to different people. We share proofs and thus share knowledge with each other. So, it seems, mathematics is at least inter-subjective unlike, say, matters of personal taste. The traditional *intuitionists*, L. E. J. Brouwer (e.g. [1912], [1948]) and Arend Heyting (1956) take a Kantian line, holding that mathematical objects arise from the mental fabric shared by all humans. Mathematics concerns the ever present *possibility* of construction.

Nominalists adopt a more thoroughgoing rejection of the objective existence of mathematical objects (e.g. Field [1980], Hellman [1989], Chihara [1990]). They insist that mathematical objects do not exist at all, in any sense. Usually, this is accompanied by a view that no abstract objects exist, perhaps with the theme that everything that exists must be subject to causal influences.

Some philosophers hold that numbers, points, functions, and sets are *properties* or *concepts*, distinguishing those from objects on metaphysical or semantic grounds. I would classify these philosophers according to what they say about properties or concepts. If such a philosopher holds that properties or concepts exist independent of language and the mind—a realism concerning properties—then she is a realist in ontology concerning mathematics. If, instead, she is a traditional nominalist about properties and concepts, then she is a nominalist about mathematics.

There is a quip, attributed to Georg Kreisel, that the important issues in the philosophy of mathematics do not concern the nature of mathematical objects, but rather the *objectivity* of mathematical discourse. This raises questions of epistemology and semantics, and concomitant issues concerning the methodology of mathematics: How do we *do* mathematics? How can mathematics be taught? Is mathematics the sort of enterprise that, when done properly, produces knowledge? That is, do we know mathematical statements, such as $7 + 5 = 12$, and that there are infinitely many prime numbers? If so, how? Clearly, the most common way to extend mathematical knowledge (if there is such a thing) is through *proof*. What is the nature of proof? How are the premises or axioms of a proof known (if they are)?

The quip from Kreisel notwithstanding, answers to these epistemic questions are surely tied to the above ontological ones. How mathematics is known should have something to do with what it is about. If a philosopher holds that mathematical objects have a given nature, say that they are abstract and acausal, then the philosopher must show us how we can know things about objects like that.

Turning more directly to semantics and logic, how is mathematical *language* to be understood? What is the underlying syntax of mathematical statements? Numerals, for example, appear to be singular terms, something like proper names. Are they? A positive

answer pushes back to the ontological matters. If numerals are, in fact, singular terms, what, if anything, do they denote?

Is mathematical language bivalent? In other words, for any given well-defined and unambiguous mathematics statement Φ, is it the case that Φ is either true or false? What is the underlying logical form of mathematical statements? And, on the heels of that question, what is the proper *logic* for mathematics? Is there a single logic that is correct for all of mathematics, or are there branches of mathematics that invoke different logics? For that matter, is there a single semantics for all of mathematics, or do the languages of different mathematical theories require different semantic treatments?

There are a number of positions on our second batch of questions, some of which are analogues of views concerning ontology. Define *realism in truth-value* to be the view that at least some mathematical statements have objective truth-values, independent of the minds, languages, conventions, etc. of mathematicians. The truth-value realist presumably holds that mathematical language is bivalent. Bivalence seems to be part and parcel of objectivity (so long as there is no vagueness or ambiguity). And most truth-value realists hold that the law of excluded middle is valid, endorsing so-called classical logic.

If mathematical statements are objective then it is possible that the truth of some of them lies beyond the abilities of humans to know these truths. So the realist in truth-value countenances the possibility that there may be unknowable mathematical truths. According to that view, truth is one thing, knowability another.

Broadly speaking, there are two varieties of irrealism in truth-value, corresponding to the two main varieties of ontological irrealism. One is that some mathematical statements have non-trivial truth-values, but not objectively. The aforementioned traditional intuitionists, Brouwer and Heyting, hold this, for much the same reasons as they hold their views on ontology. The truth-values of true and false mathematical statements are tied to the mind. The human mind is somehow constitutive of mathematical truth. Of course, it does not follow that we *decide* whether a given statement is true or false, just as an idealist about physical objects holds that we do not decide what perceptions to have.

As we saw, the truth-value realist countenances the possibility of unknowable mathematical truths. Our truth-value idealist might reject this possibility, arguing that all mathematical truths are knowable. If, in some sense, mathematical statements get their truth-values in virtue of the mind, then it would be reasonable to contend that no mathematical truth lies beyond the human ability to know. This view is sometimes dubbed *anti-realism*, or Dummettian anti-realism. If, indeed, all truths are knowable, then modesty would counsel against bivalence. It is arrogant to think that the human mind is capable of determining, of every unambiguous mathematical sentence, whether it is true or false. Some anti-realists take their view as entailing that classical logic must be replaced by intuitionistic logic (e.g. Dummett [1973], Tennant [1997]).

A second, more radical version of irrealism in truth-value is that mathematical assertions lack (non-trivial, non-vacuous) truth-values altogether. Strictly speaking, it would follow that there is no (non-trivial, non-vacuous) mathematical knowledge either so long as we agree that "Φ is known" entails "Φ is true".

Recall that nominalism is the rejection of the existence (in any sense) of mathematical objects, such as numbers, sets, and functions. One variety of nominalism is a kind of fictionalism, likening mathematical objects, such as numbers, to the objects and characters in

made-up stories (Field [1980]). On a view like this, mathematical statements are either trivially false, vacuously true, or without truth-value—depending on the fictionalist's views on the semantics of fiction. The semantic status of the sentence "7+5=12" is the same as that of "Perry Mason rarely loses a case". Other nominalists interpret mathematical language so that the non-trivial truth of its statements does not require the existence of mathematical objects. This second sort of ontological irrealist is, typically, a truth-value realist. One common route is to invoke a kind of modality in interpreting mathematics (Chihara [1990], Hellman [1989]).

It is natural for realists in truth-value to also be realists in ontology. This is just to take the language of mathematics at face-value. Or, to put it less charitably, it is to take the surface grammar of mathematical language as reflecting the logical form of the statements. Numerals appear to be singular terms; the realist says that they *are* singular terms. And numerals denote natural numbers. Since, for the truth-value realist, the statements have objective truth-values, it would seem that the numbers exist, objectively.

As a sort of contrapositive of the above, it is also natural for an ontological irrealist—idealist or nominalist—to be an irrealist in truth-value. Again, this is to take the language of mathematics at face-value (or to take surface grammar seriously). If numerals, say, are singular terms, and numbers do *not* exist, then we are dealing with statements involving non-denoting singular terms. So they are either indeterminate or vacuous.

As noted, however, some nominalists are also realists in truth-value. And one prominent philosopher, Neil Tennant (1997) combines Dummettian anti-realism in truth-value with ontological realism. So each of the four possibilities is exemplified by at least one philosopher of mathematics.

3. Benacerraf's Dilemma

Over the past forty years, a large chunk of the work in the philosophy of mathematics has been, directly or indirectly, a reaction to a dilemma posed by Paul Benacerraf, in his classic paper, "Mathematical truth" (1973). It is fair to say that this paper still dominates our discipline.

One strong desideratum is that mathematical statements should be interpreted the same way as ordinary statements, or at least as respectable scientific statements are understood. In other words, the philosopher should attempt a uniform semantics that covers both mathematical language and ordinary language, or at least the regimented languages of science. One motivation for the desideratum comes from the fact that scientific language is thoroughly intertwined with mathematical language. It is hard to make serious scientific statements that do not in some way have mathematics built in. It would be awkward and counterintuitive to provide separate semantic accounts for mathematical and scientific language, and yet another account of how the discourses interact, not to mention a semantic account for mixed statements, such as the inverse square principle concerning forces.

Someone who is a realist in both ontology and truth-value can do well on this score, assuming some sort of realism for the sciences. And, as noted, the grammar of ordinary, simple mathematical statements suggests that numerals, say, are singular terms, at least

in some uses within mathematics itself. To take a crude example, the apparent truth-conditions of the simple sentence "Seymour is human" would be that the person Seymour has the property of being human. So, from uniformity, the truth-conditions for "Seven is prime" would be that the number seven has the property of being prime. The latter is true. Ergo, numbers exist.

But we are then confronted with the second horn of Benacerraf's dilemma, that of providing a plausible epistemology for mathematics. Indeed, our combined realisms come with seemingly intractable epistemological problems. From the realism in ontology, we have the objective existence of mathematical objects. The mathematician has no need to inflect the language with tense, and there seems to be no point in pondering what caused the numbers to exist, and how they—the numbers themselves—interact with ordinary objects. So mathematical objects—if such there be—seem to be eternal, and they are outside the causal nexus. In short, numbers are abstract. But then how can we *know* anything about mathematical objects? Unless our realist posits a sort of quasi-mystical connection between the human mind and the eternal, detached, unchanging mathematical universe, he is left with a deep puzzle. How can we have any confidence in what the mathematicians say about mathematical objects? Why think that the pronouncements of mathematicians—say the sentences published in journals—are true?

Some varieties of ontological irrealism do well on the second horn of Benacerraf's dilemma: they have tractable accounts of mathematical knowledge. The traditional intuitionist follows Kant and ties mathematics to features of the mind. To put it crudely, for the intuitionist, we know mathematics to the extent that we know our own minds. The fictionalist, for her part, holds that mathematical objects do not exist. It would follow that, strictly speaking, there is no non-trivial, non-vacuous mathematical knowledge, and so nothing to account for on the second horn of the dilemma. For the fictionalist, the pursuit of mathematics involves tracking the consequences of the "stories" we tell about the fictional objects called "natural numbers", "complex valued functions", and the like. If knowledge of what follows from what is tractable, then, for the fictionalist, so is the non-trivial pursuit of mathematics—what passes for mathematical knowledge on that view.

But these irrealists come up short on the first horn of the dilemma. Unless they are idealists or fictionalists about everything (the latter being an especially strange view), then they must provide separate semantic theories for mathematical and ordinary or scientific discourse, as well as an account of how mathematics is applied in the sciences, so that we cannot so much as state a scientific proposition without invoking mathematics.

So the dilemma, in sum, is this: the desired continuity between mathematical language and everyday and scientific language suggests the two realisms, in ontology and in truth-value, but this leaves us with seemingly intractable epistemic problems. We must either solve the epistemological problems with realism, showing how we can know things about abstract objects, or give up the continuity between mathematical and everyday discourse, or give up prevailing semantical accounts of ordinary and scientific language.

4. Ontological Realism: Grasping the Epistemic Horn

Fraser MacBride (2008) glosses the aforementioned problem for the ontological realist as one "of explaining how mathematicians can reliably access truths about an abstract realm to which they cannot travel and from which they receive no signals". He writes:

> For whereas, on the face of it, the singular terms of mathematical discourse refer to inert abstract objects that are not located in space and time, mathematicians are essentially inhabitants of a concrete realm of causation and change. So a face-value or realist interpretation of mathematics appears to make a mystery of how mathematicians access truths about the mathematical domain. To access them it appears that mathematicians must do the impossible: they must transcend their own concrete natures to pass over to the abstract domain.
>
> (MacBride [2008, 156])

I think that the epistemological issues for the realist are exacerbated by the metaphorical talk of "an abstract realm" to which humans "cannot travel and from which they receive no signals". To be sure, it is part of what it is to be abstract that mathematical objects (if such there be) are not located in space and time. It does not follow from this, however, that the objects are located somewhere else, in an "abstract realm", and that to know anything about them we have to somehow "access" this realm. Abstracta like this are not located *anywhere*.

Of course, to point out that a serious problem is exacerbated is not to offer a solution to it, nor do I simply dismiss the problem. Let us ponder just what the problem is, and what would count as solving it. What, exactly, must the ontological realist establish, and on what standards? What resources can be presupposed in giving the epistemology?

There are two extreme meta-epistemological views. One of them makes the "access" problem virtually insoluble, unless the ontological realist makes some wildly implausible assumptions; the other makes it a non-problem, with a trivial solution.

The first of the extreme meta-philosophical perspectives demands some sort of reductive epistemology. When giving an account of mathematical knowledge, one cannot presuppose any mathematics—or at least one cannot presuppose that any mathematics is known. Philosophers must describe both the human mathematician and the processes they use to obtain mathematical knowledge in thoroughly non-mathematical terms, and then show that these knowers do indeed end up with mathematical knowledge. For the ontological realist, the result would be a demonstration that certain mental processes—say sequences of neuron firings—result in knowledge of a realm of abstract objects, such as numbers, points, and sets. The only straightforward route I can see for doing this is to postulate some sort of connection between the human mind and mathematics. The realist must hold that humans can somehow see that certain mathematical propositions are true, and then deduce other propositions from these. Even then, there might be an issue as to whether this direct apprehension counts as *knowledge*.

This sets too high a standard. It is widely held today that it is not possible to ground mathematics, or mathematical knowledge, in any domain that is somehow more secure than mathematics itself. It is clear that ordinary scientific scrutiny of just about anything is

going to involve mathematics, and the philosopher can make use of that mathematics like anyone else can. A reductive epistemology is too much to ask for.

The other extreme meta-epistemological view is the opposite of this. The philosopher insists that mathematics becomes known when it is obtained by the techniques sanctioned by professional mathematics (see, for example, Burgess and Rosen [1997] and Maddy [1997], [2007]). The only job for the thoroughly naturalistic epistemologist is to *describe* these techniques, an exercise in sociology, focused on the community of mathematicians. The philosopher takes it that we obviously do have mathematical knowledge—mathematicians surely know what they are doing—and leaves it at that.

Of course, this second perspective does not advance the philosophical agenda of *illuminating* the place of mathematics in the overall intellectual enterprise. It just notes that mathematics obviously has a place in our best intellectual efforts, and refuses to ponder the epistemology any further. The philosopher just relies on an entrenched (and surely plausible) assumption that mathematics, as practiced, is by and large OK as it is. What would be missing is an explanation of how the proposed "epistemology" links up with the proposed ontology. How do the techniques adopted by mathematicians and others concerning mathematics lead to knowledge *about* whatever it is that the philosopher says mathematics is about?

Arguably, philosophy is not a deductive enterprise proceeding from self-evident first principles. A philosophy of mathematics should be judged on more holistic standards, turning on how well it accounts for the ontology, epistemology, and semantics of mathematics. The ontological realist has a story to tell about the nature of mathematical objects. On the epistemology side, the burden is to make it plausible that the techniques invoked by mathematicians, both in choosing axioms and in proving theorems, do indeed result in knowledge of the sort of objects postulated in the ontology. I take this "make it plausible" in the sense of an inference to the best explanation. The ontological realist does not have to *prove that* the techniques of mathematics results in this knowledge, on some sort of non-mathematical ground. Rather, the package as a whole—ontology, epistemology, semantics—should provide a compelling, explanatory account of mathematics and its place in the intellectual enterprise.

5. Ontological Irrealism: Grasping the Semantic Horn

As the saying goes, the devil is in the details. Different ontological irrealists address the Benacerraf dilemma in different ways. As we have noted, one group of philosophers combines nominalism with realism in truth-value. The idea is to eschew a face-value interpretation of the statements of mathematics or, as they sometimes put it, to not get too hung up with the surface grammar of mathematical language. The most common route is to invoke modality, so that mathematical propositions tell us what is and what is not possible. Of course, this violates the desideratum behind the semantic horn of the dilemma, since it proposes that mathematical discourse be interpreted in a different manner than ordinary discourse (where, at least sometimes, surface grammar is taken more seriously

for semantics). The irrealist insists that the semantic accounts can be complementary, if not identical to each other.[2] Such accounts can still score well, on holistic criteria, if they can make it plausible that the ordinary techniques in mathematics, whether of choosing axioms or proving theorems, do indeed lead to knowledge of the sort postulated in the semantic account, typically modal knowledge.

These nominalists might find themselves subject to an analogue of the epistemic horn. They need an epistemology for the invoked modality, one that squares with how we have modal knowledge, whether about mathematics or anything else. In addition to that, our nominalists also need an account of how mathematics—as they understand it—is applied in science and ordinary life, showing that we can indeed illuminate the material world around us by invoking mathematics, as they understand mathematics. As part of this, they would provide a semantic account of mixed discourse, covering such things as the inverse square principle and what is involved in counting one's fingers and balancing a checkbook.

The fictionalist need not confront the semantic horn of the Benacerraf dilemma. She can follow the ontological realist in adopting a face-value reading of mathematical discourse—just as she can adopt a face-value reading of fictional discourse. And, as above, there need not be a serious problem on the other, semantic, horn as well. For the fictionalist, there is no mathematical knowledge, strictly speaking. Instead, mathematics is the discovery of what follows from the "stories" behind each theory. Of course, the fictionalist will also hold that mathematical statements, strictly speaking, are either indeterminate or vacuous. Her main problem lies in explaining the role of mathematics in science and everyday life. Why are the fictional stories concocted by mathematicians so useful—indeed essential—in understanding just about anything?

How well the idealists do on our dilemma may depend on how far the idealism extends. If the idealism does not extend to ordinary objects and the like, then the semantic horn of the Benacerraf dilemma turns on whether statements about ordinary objects should have the same sort of semantics as those that speak of mental constructions—mathematical objects in the case at hand. If the idealism does extend to ordinary objects, then the semantic horn of the dilemma seems to be met—assuming there is a plausible semantics for ordinary discourse, from the perspective of the idealist. This is not the place to reiterate the issues concerning a global idealism.

6. Mathematics and Philosophy of Mathematics

There is a longstanding perspective that philosophy *precedes* practice in a metaphysical sense. At a fundamental level, philosophy determines practice. Concerning the present topic, the idea is that the philosopher describes or discovers what mathematics is all about—whether, for example, mathematical entities are objective or mind-dependent. This

[2] As Tennant (1997) notes, the same goes for the combination of ontological realism and truth-value irrealism.

fixes the way mathematics is to be done. I have called this the *philosophy-first principle* (e.g. Shapiro [2000, Chapter 1]). The first-philosopher has it that, on a metaphysical level, we first figure out what it is that we are talking about and only then figure out what to say about it—in mathematics itself. In general, philosophy supplies first principles for the special sciences like mathematics.

Plato is the quintessential first-philosopher. As noted above, he held that the subject matter of mathematics is an eternal, unchanging, realm. Its objects are not created and destroyed, nor can they be changed. However, virtually every source of ancient geometry (and, indeed, much of the current literature on that subject) makes use of constructive, dynamic language: lines are drawn, figures are moved around, etc. In Book VII of *Republic*, Plato complained that geometers do not know what they are talking about, and for this reason they do mathematics incorrectly:

> [The] science [of geometry] is in direct contradiction with the language employed by its adepts . . . Their language is most ludicrous . . . for they speak as if they were doing something and as if all their words were directed toward action . . . [They talk] of squaring and applying and adding and the like . . . whereas in fact the real object of the entire subject is . . . knowledge . . . of what eternally exists, not of anything that comes to be this or that at some time and ceases to be.

Clearly, one cannot take an eternal, unchanging line segment and cut it in half and then move one of the parts on top of another figure. For Plato, if his philosophy is the correct one, then mathematics has to change, at least at the level of its language.

Closer to our time, debates over intuitionism provide another philosophical challenge to mathematics as practiced. As noted above, the traditional intuitionists, Brouwer and Heyting, held that mathematical objects are mental constructions, and mathematical statements somehow invoke mental construction. Because of this, they deny bivalence and the validity of the so-called law of excluded middle, in symbols $\Phi \vee \neg\Phi$. Brouwer and Heyting argue that excluded middle, and related principles based on it, are symptomatic of faith in the transcendental existence of mathematical objects and/or the transcendental truth of mathematical statements.

Later in the twentieth century, Dummett (e.g. [1973]) brought a host of considerations concerning the learnability of language and the use of language as a vehicle of communication. One consequence is that the law of excluded middle is not generally valid and that classical logic should be replaced by intuitionistic logic. Dummett, of course, is aware that if he is right about language then contemporary mathematical practice is flawed—and indeed incoherent. Those inclined toward philosophy-first might take Dummett's arguments concerning language seriously. It is a live possibility that Dummett is right and that just about every mathematician is incoherent, or at least badly mistaken on a regular and systematic basis.

There are other methodological disputes within mathematics that are (or were) sometimes construed as turning on philosophical considerations. A definition of a mathematical entity is said to be *impredicative* if it refers to a collection that contains the defined entity. For example, the usual definition of "least upper bound" is impredicative since it refers to a set of upper bounds and characterizes a member of this set. Henri Poincaré based a systematic attack on the legitimacy of impredicative definitions on the idea that mathematical objects do not exist independently of the mathematician (e.g. Poincaré [1906], see

Goldfarb [1988]). The set of real numbers is not fixed prior to the mathematical activity. From this perspective, impredicative definitions are viciously circular. One cannot construct an object by using a collection that *already* contains it.

Against this, Gödel (1944) made an explicit defense of impredicative definition, based on his philosophical views concerning the existence of mathematical objects:

> the vicious circle . . . applies only if the entities are constructed by ourselves. In this case, there must clearly exist a definition . . . which does not refer to a totality to which the object defined belongs, because the construction of a thing can certainly not be based on a totality of things to which the thing to be constructed belongs. If, however, it is a question of objects that exist independently of our constructions, there is nothing in the least absurd in the existence of totalities containing members, which can be described (i.e., uniquely characterized) only by reference to this totality . . . Classes and concepts may . . . be conceived as real objects . . . existing independently of us and our definitions and constructions. It seems to me that the assumption of such objects is quite as legitimate as the assumption of physical bodies and there is quite as much reason to believe in their existence.

According to Gödel's ontological realism, a mathematical "definition" is not a recipe for constructing, or otherwise creating, an object. Rather, it is a way to characterize or point to an already existing thing. Thus, an impredicative definition is not viciously circular. "The least upper bound" is no more problematic than other "impredicative" definitions, such as the use of "the star player of the game" to denote the player who did the best in the given game.

I think it is fair to say that most contemporary philosophers do not subscribe to the philosophy-first principle.[3] The prevailing attitude, I think, is that the target of our philosophical reflection is *mathematics*—mathematics as practiced—and not what a prior philosophy takes mathematics to be. Of course, mathematicians are human and are subject to mistakes, and it is conceivable that some mistakes within mathematics can be found, and corrected, through something recognizable as philosophy, but it is taken for granted, at least as a strong working hypothesis, that mathematicians know what they are doing, and that the vast majority of what they are doing is correct.

Philosophers have their own interests, beyond those of their colleagues in other departments, and the pursuit of those interests is interesting and worthwhile. The work of the philosopher of mathematics should merge with that of the mathematician, but at least part of it is different work. Philosophy and mathematics are intimately interrelated, with neither one dominating the other. On this view, the correct way to do mathematics is not a direct consequence of the true philosophy, nor is the correct philosophy of mathematics an immediate consequence of mathematics as practiced.

As noted, Dummett argues that, based on considerations of what language is and how it functions, much of contemporary mathematics does not enjoy a certain level of justification. A philosopher can agree with this, and then wonder whether mathematics needs this level of justification. Or, to sound a holistic theme again, we can wonder if Dummett's

[3] I might add that the philosophy-first principle is not a dominant theme in Gödel's own published philosophical papers. The purpose of Gödel (1944) is to *respond* to a philosophically based attack on mathematical principles. His argument is that the methodological criticisms are based on a philosophy that one need not adopt.

account of language and how it functions is sufficiently compelling that we can demand revisions in mathematics as practiced. Which is more secure and more likely to be correct, mathematics as practiced or Dummett's philosophy of language?

7. Naturalism and Mathematics

Issues concerning naturalism follow closely on the heels of the rejection of philosophy-first.[4] W. V. O. Quine (1981, 72) characterizes *naturalism* as "the abandonment of the goal of first philosophy" and "the recognition that it is within science itself … that reality is to be identified and described". The Quinean naturalist contends that epistemology, and philosophy generally, must be continuous with science, ultimately physics: "The naturalistic philosopher begins his reasoning within the inherited world theory as a going concern … [The] inherited world theory is primarily a scientific one, the current product of the scientific enterprise." The naturalist looks at physical science "as an inquiry into reality, fallible and corrigible, but not answerable to any supra-scientific tribunal, and not in need of any justification beyond observation and the hypothetico-deductive method" (Quine [1981, 72]).

The epigraph to Quine's influential (1960) is a quotation from Otto Neurath (1932), which can be translated as follows: "We are like sailors who have to rebuild their ship on the open sea, without being able to dismantle it in dry dock and reconstruct it from the best components." The philosopher is just as much a "sailor" as the physicist, the chemist, and the psychologist. Quine does not include the next sentence in Neurath's text, which is "Only metaphysics can disappear without trace." At least part of metaphysics is an integral part of the "ship".

So what does the naturalist make of mathematics? There is an interesting irony in Quine's focus on *science*, especially on physics. For Quine, the driving goal of the scientific-cum-philosophical enterprise—the ship of Neurath—is to account for and predict sensory experience. Quine thus accepts *mathematics* only to the extent that it is needed for the scientific/philosophical enterprise (perhaps with a little more mathematics thrown in, for "rounding things out"). In particular, he does not accept (as true) the parts of mathematics, such as advanced set theory, that go beyond this role of aiding and abetting empirical science. In short, for Quine, if a part of mathematics does not play an inferential role (however indirect) in the parts of the scientific-web that bear on sensory perception, then that part should be jettisoned, via Occam's razor.

Quine's proposal here smacks of something in the neighborhood of philosophy-first, even if it is science/philosophy-first. Quine makes proposals *to mathematicians*, based on his overall philosophy of mathematics and science. He suggests, for example, that set theorists adopt a certain principle (called "V=L") on the ground that the resulting theory is clean, and so presumably easier to apply. The fact that most (but not all) practicing theorists are skeptical of this principle is not relevant to Quine's views, no more than the fact that most mathematicians adopt excluded middle is relevant to the intuitionistic critique.

[4] The bulk of this section is a summary of Shapiro (2000, Chapter 1, §3).

Penelope Maddy's (1997), (2007) version of naturalism prescribes a deferential attitude towards mathematicians much like the one Quine shows toward scientists. Mathematics has its own methodology, which has proven successful over the centuries. The success of mathematics is measured in mathematical, not scientific terms.

Against Quine, one might argue that if mathematicians gave serious pursuit only to those branches *known to* have applications in natural science, we would not have much of the mathematics we have today. More importantly, we would not have all of the *science* we have today. The history of science is full of cases where branches of "pure" mathematics eventually found application in science. In other words, the overall goals of the scientific enterprise have been well served by mathematicians pursuing their own discipline *with their own methodology* and, in particular, without any focus on applications in science.

This argument should have force within Quine's overall holistic empiricist framework. In a sense, Maddy argues that Quine mis-applies his own methodology for philosophy. We do not need even an indirect inferential link between a piece of mathematics and sensory experience before we can accept mathematics as a legitimate part of the enterprise.[5]

8. How Much Mathematics?

It seems to be a truism that philosophers should know something about whatever it is that they want to philosophize about. So there is a question concerning how much mathematics the philosopher should be familiar with, and which branches of mathematics and which theorems the philosophy should concern. Of course, the easy answer to this question is "all of it": the philosopher should be familiar with the length and breadth of mathematics—all of its branches and all of its results.

Clearly, that would be asking too much. There are few professional mathematicians who are familiar with branches and results removed from his or her own specialty. Each branch has its own techniques, some of which take considerable time to digest and deploy.

At one extreme are philosophies that limit their attention to basic arithmetic and do not know much mathematics beyond what is taught in high schools—if that much. This practice is not entirely unfair and worthless, provided only that the philosopher recognizes its limitations. Mathematics and its most basic objects, such as natural numbers and some real numbers (or at least rational numbers), play an important role in everyday life, well beyond the focus of professional mathematicians and scientists. We count objects, and balance checkbooks; we measure land, flour, gasoline, and the like. There are indeed interesting and important philosophical questions about these activities, and the objects they involve (if there are any such objects). One can address such questions without pondering any advanced mathematics.

[5] It should be added that Maddy does not endorse Quine's overarching holism, and so does not accept the metaphor of a web of belief—the "ship of Neurath". She takes the seams in the web of belief seriously, and holds that we do not have to show that there is an ultimate deductive connection to science in order to justify a piece of mathematics. There are legitimate goals beyond the prediction and control of sensory experience.

One interesting batch of questions concerns the role of cardinal and ordinal numerals in ordinary language. This is an important topic within linguistics, but it has a philosophical pedigree as well. Frege noted that cardinal terms appear in natural language as both adjectives, as in "we own four dogs", and as singular terms—proper nouns—as in "the number of dogs we own is four". Frege held that the singular-term use is the primary one, with the adjectival use derivative. According to Dummett (1991), this linguistic conclusion drives Frege's views that numbers are objects—his realism in ontology. That is, Frege believed that numbers are objects *because* he believed that numerals are singular terms. It was not the other way around. To this extent, then, for Frege, linguistics determines metaphysics.

But perhaps Frege was mistaken in his linguistic conclusions. A group of philosophers examine the use of numerical terms in natural language, to see what philosophical conclusions should be drawn, if any. Often, this takes the philosopher to delicate issues in the philosophy of language and even to linguistics.

So, I suggest that the focus on elementary mathematics is legitimate. Of course, we'd also like to see how this everyday mathematics relates to the activities of professional mathematics, and to the mathematics used in science and engineering.

Beyond the everyday, the most common focus of philosophers of mathematics is on natural numbers, real numbers, complex numbers, various functions, geometry, and the more foundational branches of mathematics, such as set theory and perhaps category theory. It is not uncommon for philosophers to have formal training in advanced mathematics, often focused on the foundational branches.

There are also philosophers who focus on the more advanced areas of general mathematics—not just the foundational bits. Some issues demand this (see sections 9.1–11.1 for some examples). There is a recently formed society for the study of mathematics as practiced. It is also not uncommon for philosophers to be familiar with various episodes in the history of mathematics and science. I'll end this survey with a brief account of a few topics within the philosophy of mathematics that have their own methodological issues.

9. Explanation

One of the most vexed areas in philosophy concerns the nature of explanation. Intuitively, an explanation is an answer to a "why" question. As Aristotle put it, "men do not think they know a thing till they have grasped the 'why' of it" (*Physics*, Chapter 3). The problem is to figure out when, in fact, someone has grasped the "why" of something. The one area that has paid the most attention to explanation is the philosophy of science, although, naturally, most of that work is focused on scientific explanation. Mathematics makes for an especially compelling case. There are two different areas of focus.

First, there are cases where, it seems, a mathematical fact is cited as an explanation of non-mathematical phenomenon (or at least as a crucial aspect of such an explanation). Suppose one wants to know why rain forms into drops. A typical explanation starts with the basics of surface tension, and then cites the fact that a sphere is the largest solid with a given surface area (or that a sphere has the smallest surface area for a given volume).

If things like this are genuine explanations, they serve as counter-examples to some once widely held accounts of scientific explanation (see, for example, Salmon [1989]). For some time, the received view was that an explanation consists of deriving the explanandum from a Law of Nature, plus some relevant initial conditions (Hempel and Oppenheim [1948]). That would rule out the above example, as an explanation, unless one somehow stretches the notion of a Law of Nature to include the mathematical fact about spheres (or chalks the deep mathematical fact to the "deduction" of the phenomenon). Another popular account takes an explanation to be the giving of a relevant part of the causal history of the explanandum. Thus, one explains the tides by showing what causes them (e.g. Salmon [1984]). This also rules out the foregoing example. Mathematical facts, such as the above proposition about spheres, do not cause anything, or so it seems. In particular, the fact that spheres provide the maximal volume for a given surface area does not seem to "cause" moisture to form into drops.

Of course, this is not the place to delve into this most interesting philosophical problem. Dealing with it adequately requires a good working knowledge of both the mathematics and the underlying science that goes into such explanations. For all but the most simple cases, the philosopher must go beyond the foundational areas of mathematics. Those are rarely cited in explanations.

The second sort of cases are explanations within mathematics itself. In particular, some proofs are regarded as explanatory and some not. Of course, any legitimate proof, in any area of mathematics, establishes *that* its conclusion is true (or correct, or whatever). An explanatory proof also gives insight into *why* the conclusion is true. It seems that there is a wide consensus on the various cases: mathematicians usually agree, by and large, on which proofs are explanatory and which are not.

The phenomenon of explanatory proof is further evidence against the aforementioned accounts of (scientific) explanation. The notion of a Law of Nature does not seem to apply in mathematics, at least not as a distinction among generalizations, and talk of causality is surely out of place, at least in pure mathematics. Of course, one might dismiss the enterprise by arguing that explanatory proofs are not genuine explanations, but pending some argument for this, it seems like ad hoc monster barring.

Here, too, we have an interesting and potentially important issue in the philosophy of mathematics that requires a detailed familiarity with a wide body of mathematics, well beyond the foundational areas. If nothing else, philosophers need an extensive body of examples of explanatory and non-explanatory proofs to work with. They also need to be familiar with the extensive literature on explanation—scientific and otherwise. That is what it takes to be able to say something illuminating about this most interesting phenomenon.

To be sure, it is possible that there is nothing interesting and illuminating that characterizes (and illuminates) all and only explanatory proofs in mathematics. However, such a pessimistic conclusion should only come after one has attempted a sufficiently detailed analysis of the various cases. I think it fair that this is, so far, an under-explored area (see Steiner [1978], [1980], and the contributions in Mancosu, Jørgensen, and Pedersen [2005], especially Hafner and Mancosu [2005]).

10. Applications of Mathematics

The first batch of cases presented in the previous section—where mathematics is cited in explanations of non-mathematical phenomena—points toward a more general group of problems concerning the relationship between mathematics and physical (and psychological and social) reality. Why is it that mathematics seems to be so important in the scientific study of just about anything? What does that say about the nature of mathematics, and about the nature of non-mathematical reality?

Mark Steiner (1995), (1997) distinguishes several philosophical problems that fall under the rubric of "applying mathematics". There are, first, *semantic* matters. Typical descriptions of non-mathematical phenomenon—scientific or otherwise—invoke a mixture of mathematical and physical terms. This goes for simple statements like "I have three children" and "My bank account has a negative balance", and, of course, just about any sophisticated theory within modern science. The problem is to find an interpretation of the language that covers "pure" and "mixed" contexts, so that proofs within mathematics can be employed directly in scientific contexts. This is of-a-piece with the first horn of the aforementioned Benacerraf dilemma (§3).

A fully adequate treatment of such matters would require familiarity with semantics, and thus takes the philosopher to issues in the philosophy of language and even in linguistics. Indeed, a large part of the problem is to give an account of mathematical and ordinary language. Surely, it would help to know something about languages generally. Note that this also dovetails with issues, broached briefly in §8, concerning the semantics of cardinal and ordinal terms in ordinary language.

A second group of problems are in the general realm of metaphysics. How do the objects of mathematics (if such there be) relate to the physical world, so that applications are possible? On a typical ontological realism, for example, mathematics is about a realm of causally inert abstract objects. On a typical idealism, mathematics is about mental activity. In either case, how can stuff like that tell us anything about how the physical world works? Of course, this recapitulates the second horn of the Benacerraf dilemma.

A third group of issues concerns why the *specific* concepts and formalisms of various branches of mathematics are useful in describing empirical reality. These matters can be raised for elementary and advanced branches of mathematics alike. What is it about the physical world that makes arithmetic so applicable? What is it about the physical world that makes group theory and Hilbert spaces central to describing it? Steiner suggests that we really have a different problem here for each mathematical concept that finds application, and so one should not expect a uniform solution. To take a crude example, the applicability of arithmetic in ordinary discourse depends on the medium-sized macroscopic world being naturally (or at least seemingly) divided into stable objects (and organisms) that keep their sizes and shapes over at least short periods of time. The applicability of real analysis depends on the existence of events and processes that are at least approximately continuous.

Steiner delimits a compelling, related group of problems. Occasionally, areas of pure mathematics, such as abstract algebra and analysis, find unexpected applications long after

their mathematical maturity. Mathematicians seem to have an uncanny ability to come up with structures, concepts, and disciplines that find unexpected application in science.

Throughout history, the following scene played itself out repeatedly. Mathematicians study a given structure, for whatever reason. They extend it to another structure for their own, internal purposes (say by considering infinitely many dimensions); and then later the newly defined structure finds application somewhere in science. As S. Weinberg (1986, 725) put it: "It is positively spooky how the physicist finds the mathematician has been there before him or her". A similar sentiment was echoed by the Bourbaki conglomeration (1950, 231): " … mathematics appears … as a storehouse of abstract forms—the mathematical structures; and it so happens—without our knowing why—that certain aspects of empirical reality fit themselves into these forms, as if through a kind of preadaption".

Of course, even if it looks this way, it is not easy to establish that it *is* this way. Is it really the case that a discipline that begins life as a chunk of pure mathematics finds application significantly often? What does this mean? How could we tell? And if so, then why is this? Clearly, there are interesting issues here to confront. To adequately deal with them, the philosopher must be familiar with a large body of mathematics and a large body of science.

11. Purity

Throughout history, mathematicians have often expressed a preference for solutions to problems that do not invoke concepts and methods that are somehow "foreign" to the problem under investigation. For example, a solution to a problem in, say, arithmetic or real analysis should not invoke geometric matters. It is not clear, a priori, just what this "purity" comes to, nor why it is valued. Perhaps there are different kinds of purity, with differing levels of value (see, for example, Detlefsen and Arana [2011]).

There is also a tendency in the opposite direction. Sometimes, it seems, an area of mathematics is best illuminated when it is embedded in another, seemingly "foreign environment". To invoke a passage from Kreisel's "Informal rigour and completeness proofs" (1967) that I have quoted elsewhere (more than once):

> very often the mathematical properties of a domain D become only graspable when one embeds D in a larger domain D′. Examples: (1) D integers, D′ complex plane; use of analytic number theory. (2) D integers, D′ p-adic numbers; use of p-adic analysis. (3) D surface of a sphere, D′ 3-dimensional space; use of 3-dimensional geometry. Non-standard analysis [also applies] here.
>
> (Kreisel [1967, 166])

Invoking the complex plane to illuminate the integers at least appears to bring in something "foreign" to the integers. Thus the procedure is "impure". Yet clearly valuable.

One might claim that the "good" embeddings tell us what the original theory is "all about" in the first place. That is, we learned a surprising fact that the complex plane is not really "foreign" to the integers. A youthful and perhaps overly exuberant Hermann Weyl (1955, VII) once wrote that Riemann's approach to complex analysis should be seen

> not merely a device for visualizing the many-valuedness of analytic functions, but rather an indispensable essential component of the theory; not a supplement, more or less artificially distilled from the functions, but their native land, the only soil in which the functions grow and thrive.

See also Wilson (1992, 111).

Clearly, sorting out and coming to understand what "purity" amounts to, and why it is valued, at least sometimes, takes us into some deep issues concerning the ontology and metaphysics of mathematics. To make progress on this, one would have to be familiar with a wide range of mathematics, but, this time, in its historical context. Here, too, the philosopher who focuses only on the more foundational side of mathematics will miss something important.

12. Closing

As noted at the outset, I do not claim that this article covers, or even mentions, everything of interest within the philosophy of mathematics. And so I do not claim comprehensive coverage of the methodology of our discipline. I do hope, however, that the preceding survey provides a decent sample of the wares, and of the background skills and methodology employed by the best practitioners of our subject.

References

Aristotle, *Physics, The Basic Works of Aristotle*, R. McKeon, ed. (Randon House, 1941), pp. 316ff.

Benacerraf, P. (1973), "Mathematical truth", *Journal of Philosophy 70*, 661–679; reprinted in Benacerraf and Putnam (1983), 403–20.

Benacerraf, P., and H. Putnam (1983), *Philosophy of Mathematics*, second edition, Cambridge, Cambridge University Press.

Bourbaki, N. (1950), "The architecture of mathematics", *American Mathematical Monthly 57*, 221–32.

Brouwer, L. E. J. (1912), *Intuitionisme en Formalisme*, Gronigen, Noordhoof; translated as "Intuitionism and Formalism", Benacerraf and Putnam (1983), 77–89.

Brouwer, L. E. J. (1948), "Consciousness, philosophy and mathematics", in Benacerraf and Putnam (1983), 90–6.

Burgess, J. and G. Rosen (1997), *A Subject With No Object: Strategies for Nominalistic Interpretation of Mathematics*, Oxford, Oxford University Press.

Chihara, C. (1990), *Constructibility and Mathematical Existence*, Oxford, Oxford University Press.

Detlefsen, M., and Andrew Arana (2011), "Purity of methods", *Philosopher's Imprint 11*(2), 1–20. Available from <http://quod.lib.umich.edu/cgi/p/pod/dod-idx/purity-of-methods.pdf?c=phimp;idno=3521354.0011.002>. Accessed September 26, 2015.

Dummett, M. (1973), "The philosophical basis of intuitionistic logic", in *Truth and other enigmas*, by M. Dummett, Cambridge Massachusetts, Harvard University Press, 1978, 215–247; reprinted in Benacerraf and Putnam (1983), 97–129.

Dummett, M. (1991), *Frege: Philosophy of Mathematics*, Cambridge, Massachusetts, Harvard University Press.
Field, H. (1980), *Science Without Numbers*, Princeton, Princeton University Press.
Gödel, K. (1944), "Russell's mathematical logic", in Benacerraf and Putnam (1983), 447–69.
Gödel, K. (1964), "What is Cantor's continuum problem?", in Benacerraf and Putnam (1983), 470–85.
Goldfarb, W. (1988), "Poincaré against the logicists", in *History and philosophy of modern mathematics*, edited by W. Aspray and P. Kitcher, Minneapolis, Minnesota Studies in the Philosophy of Science, Volume 11, University of Minnesota Press, 61–81.
Hafner, J., and P. Mancosu (2005), "Varieties of mathematical explanation", in Mancosu, Jørgensen, and Pedersen (2005), 215–50.
Hellman, G. (1989), *Mathematics Without Numbers*, Oxford, Oxford University Press.
Hempel, C., and P. Oppenheim (1948), "Studies in the logic of explanation", *Philosophy of Science 15*, 135–75.
Heyting, A. (1956), *Intuitionism: An Introduction*, Amsterdam, North Holland.
Kreisel, G. (1967), "Informal rigour and completeness proofs", *Problems in the philosophy of mathematics*, edited by I. Lakatos, Amsterdam, North Holland, 138–86.
MacBride, F. (2008), "Can ante rem structuralism solve the access problem?", *Philosophical Quarterly 58*, 155–64.
Maddy, P. (1990), *Realism in Mathematics*, Oxford, Oxford University Press.
Maddy, P. (1997), *Naturalism in Mathematics*, Oxford, Oxford University Press.
Maddy, P. (2007), *Second Philosophy: A Naturalistic Method*, Oxford, Oxford University Press.
Mancosu, Paolo, Klaus Jørgensen, and Stig Pedersen (editors) (2005), *Visualization, Explanation and Reasoning Styles in Mathematics, Synthese Library 327*, Dordrecht, Springer.
Neurath, O. (1932), "Protokollsätze", *Erkenntnis 3*, 204–14.
Poincaré, H. (1906), "Les mathématiques et la logique", *Revue de Métaphysique et de Morale 14*, 294–317.
Quine, W. V. O. (1960), *Word and Object*, Cambridge, Massachusetts, The MIT Press.
Quine, W. V. O. (1981), *Theories and Things*, Cambridge, Massachusetts, Harvard University Press.
Salmon, W. (1984), *Scientific Explanation and the Causal Structure of the World*, Princeton, Princeton University Press.
Salmon, W. (1989), *Four Decades of Scientific Explanation*, Pittsburgh, University of Pittsburgh Press.
Shapiro, S. (1997), *Philosophy of Mathematics: Structure and Ontology*, New York, Oxford University Press.
Shapiro, S. (2000), *Thinking about Mathematics: The Philosophy of Mathematics*, Oxford, Oxford University Press.
Steiner, M. (1978), "Mathematical explanation and scientific knowledge", *Noûs 12*, 17–28.
Steiner, M. (1980), "Mathematical explanation", *Philosophical Studies 34*, 135–52.
Steiner, M. (1995), "The applicabilities of mathematics", *Philosophia Mathematica (III) 3*, 129–56.
Steiner, M. (1997), *The Applicability of Mathematics as a Philosophical Problem*, Cambridge, Massachusetts, Harvard University Press.
Tennant, N. (1997), *The Taming of the True*, Oxford, Oxford University Press.

Weinberg, S. (1986), "Lecture on the applicability of mathematics", *Notices of the American Mathematical Society 33*, 725–33.
Weyl, H. (1955), *The Concept of a Riemann Surface*, Reading, Massachusetts, Addison-Wesley; English translation of 3rd revised edition of Weyl *Die Idee Der Riemannschen Fläche*, Teubner, Leipzig, 1913, G. MacLane, translator.
Wilson, M. (1992), "Frege: The royal road from geometry", *Noûs 26*, 149–80.

CHAPTER 32

METHODS IN THE PHILOSOPHY OF LITERATURE AND FILM

GREGORY CURRIE

1. Introduction

Work in the philosophy of literature and film (PLF) crosses a number of boundaries.[1] Its questions lie partly within aesthetics, especially when artistic value, beauty, and the nature of criticism are in focus. But the philosophical analysis of literature also requires that we take a view on complex issues to do with meaning and communication which have their home in linguistics and the philosophy of language; film of course is a communicative device and attracts an overlapping discussion, partly because it has been said to embody a language of its own. Special features of the cinematic medium, such as its reliance on a mechanism of recording, raise metaphysical questions about the relation of film to reality and about our perceptual contact with the real world. Then there are questions about the morality of certain narrative representations, and psychological questions about our response to such representations and about the ways that complex narratives of character and incident may (or may not) educate and refine us. Thus work in PLF is bound to exhibit commonalities of method with other areas of philosophy, and stands in similarly complex and contested relations to work in other disciplines, especially empirical ones—an issue I take up in some detail later in this chapter. First I provide some background material concerning the domain and ambitions of our subject.

[1] Thanks to Anna Ichino for many discussions of these issues; we expand on the problems discussed in the final section in a joint paper, in preparation.

2. Delineating the Philosophy of Literature and Film

We should distinguish between the activity here called the philosophy of literature and film (PLF), and those philosophically influenced projects for understanding, interpreting, and theorizing these forms which I will call, collectively, the theory of literature and film (TLF) and which are more often pursued by scholars in language and literature departments or in film studies. The contrast here is one aspect of the familiar divide between analytical and continental approaches to philosophy. This border is porous, in many places hardly visible, and regarded by many as the legacy of an insular past. But within the study of literature and film it seems to be well maintained: it is common to find philosophical approaches to literature and film of both kinds exemplified in the teaching and research of an institution, one in the philosophy department and the other in media and literature departments, without significant intellectual exchange between the academics involved. The aim of this essay is to describe and, in small ways, to assess the methods used in PLF; approaches from within TLF will be brought into focus only when they help us with that task.

Unsurprisingly, there are some exceptions to this alignment between departmental and intellectual affiliations: there are philosophers whose work is hard to classify or, while comfortably on one side of the border, of interest to those on the other side.[2] And there are literary scholars whose orientation is closer to work in PLF than to TLF. This is especially noticeable with work on evolutionary approaches to literature, which some literary scholars have taken up with enthusiasm; I discuss this approach further on. The relation of this work to PLF is complicated because there is no consensus there about how, if at all, scientific work is relevant to understanding and appreciating the narrative arts. But there is a naturalistically inclined strand to thinking within PLF which is lacking in TLF, and the advocates of that strand are often friendly to Darwinian ideas.

Three other preliminary observations may be helpful, the first on terminology. "Literature" here covers drama and poetry as well as prose writing, though in practice philosophical work in this field tends to focus on the novel, with some attention to drama, frequently Shakespearean. And "film" is slightly misleading since the locus of philosophical interest is often on the broader class of representational media which work by the recording of events, and there is a strong overlap between the issues that arise when we consider film and those we confront with still photography; television and radio are also candidates for reflection under this heading, though there has in fact been little philosophical work done on the first and none on the second.

The second point is that there is an asymmetry between the philosophy of film and the philosophy of literature in that the latter (as its name suggests) is concerned with writing of a certain quality.[3] The philosophy of film on the other hand is conceived first of all as the study of a certain medium, which may be exemplified in works of any quality at all.

[2] See e.g. the work of Stanley Cavell (1971, 1981, 1987) and Steven Mulhall (2002).

[3] The term "literature" is sometimes used in a way which carries no implication of quality but "the philosophy of literature" generally does, especially when contrasted with the philosophy of fiction.

And while much in the philosophy of film is devoted to the analysis of works perceived as meriting aesthetic and artistic attention, the attractions of popular genres and downright awful films are included as well. By contrast, literature is not a medium, but an evaluative category; the medium of literature is language, and most uses of language do not produce literature.[4] Nor is there any point in trying to identify a medium which consists of "literary-language". There are tropes of language generally recognized as signs (not infallible ones) that we are in the domain of literature, free indirect discourse being one. But much of the language used to literary purpose is not literary in anything other than the sense of being available for literary purposes. No bits of language are off-limits to literature.

Thirdly, studies within PLF of literature and of film differ somewhat in the emphasis they give to the idea of fiction. While philosophers of literature accept that there is non-fictional literature (the Bible is an often-cited example), their investigations are almost always confined to fictional literature—roughly, those literary works which tell a story without any systematic commitment on the author's part to the truth of what is told.[5, 6] But philosophical studies of film sometimes focus on how to characterize the concept *documentary*, and the close analysis of particular documentary—and hence non-fiction—films plays a role in arguments about this.[7]

3. Recent History

The institutional divide between PLF and TLF, regrettable as it is in many ways, is the reflection of an intellectual division which is partly one of method, and therefore of interest to us. Take first the case of film. By the 1970s the academic study of cinema was a going concern, drawing heavily on ideas from Marx and Freud and their followers, and from semiotics. By the 1980s a group of philosophers was seeking an alternative to these theoretical commitments and to what they saw as the sometimes obscure and dogmatic style of argument that went with them. The semiotic attempt at a general theory of signs was particularly thought to be a misguided project, grounded in work on language and meaning long superseded by the impressive developments in semantics, syntax, and pragmatics which have been so much a collaborative effort of philosophers and linguists since the 1960s.[8] Two books by Noel Carroll were important here: *Philosophical Problems of Classical Cinema* was an attempt to engage, carefully and sympathetically, with an

[4] But see Carroll (2004). For an account of literature based on the idea that literary works require a certain kind of learned, and indeed institutionalized, attention, see Lamarque (2008).

[5] This is a very rough characterization of the fictional, and a disputed one: see Friend (2012) for an opposing view. I say "without *systematic* commitment" because it is often the case that the story told is understood to be presented against a reliable factual background.

[6] For example, Peter Lamarque's recent book on the philosophy of literature takes its examples almost entirely from the fictional domain, including—in accordance with the understanding of "fiction" given in the text to the preceding note—a good deal of poetry.

[7] For opposing views see N. Carroll (1997) and Currie (1999).

[8] For a critical view of semiotics, see Harman (1979); for criticism of structuralist and post-structuralist approaches to language, see Devitt and Sterelny (1999, ch.13) and, with reference specifically to the study of literature, Lamarque and Olsen (1994).

older tradition of thinking about the nature of the film medium and its implications for film style, best represented by the work of Arnheim and Bazin, while *Mystifying Movies* was a plain-speaking assault on such dogmas of TLF as the analogy between the screen and the breast, the illusion of reality in film, and the passivity of the viewer. With film scholar David Bordwell, Carroll did much to create an atmosphere in which it seemed at last possible to make headway in thinking about cinema while adhering to clarity, careful argument and, where appropriate, appeal to psychological theories for which there is some evidence.[9]

The study of literature, meanwhile, had undergone convulsive changes that ushered in post-structuralism, deconstruction, and postmodernism, combining and sometimes contesting with Marxist and Freudian ideas, which for some time had enjoyed currency in the field. Here, analytically inclined philosophers can claim membership in something approaching a rival tradition, with antecedents in Hume and Johnson.[10] These philosophers are often close in doctrine and approach to the liberal humanist thinkers once dominant, but now less common, in literature departments, Martha Nussbaum's approach to value in literature being close to that of, say, Lionel Trilling fifty years earlier, though more systematic in the defence of general claims.[11] This philosophical version of the humanistic tradition will be examined in section 7.

4. General Arguments and Particular Works

A central methodological question for practitioners of PLF is how we should understand the relations between general arguments and the analysis of particular works of literary and filmic art. Practitioners of PLF are interested primarily in general questions. Those whose focus is film wish to understand the nature of the filmic medium, its relation to still photography, its distinctness from both linguistic representation and from the "handmade" depictions of painting and drawing, and (as already described) the extent to which a viable distinction between fictional and documentary films can be made.[12] There has also been a good deal of debate—also general—on the question of how to characterize the viewer's relation to the work: is it, for example, a relationship in which the viewer imagines seeing the fictional events depicted on screen? If so, does the imaginative project involve the idea that one is seeing those events directly, or via the mediation of a visual recording of those events? Might the imagining involved not be imagining seeing, but imagining of some other kind? Competing answers to all these questions have been defended in recent years.[13]

[9] See the collection of essays edited by Bordwell and Carroll (1996).

[10] See Hume (1777/1985); Johnson (1765).

[11] See Nussbaum (1990), and Trilling (1950). While Nussbaum may be counted as closer in outlook to philosophers in the broadly analytical tradition than to, say, deconstructive thinkers, she is not straightforwardly an analytical philosopher.

[12] On transparency, see e.g. Walton (1984); Currie (1995, ch.2).

[13] See Currie (1995, ch.5–6); Walton (1997); Wilson (2011).

Much work in the philosophy of literature also falls within the scope of the generalizing project—some of it, for reasons we have given, concerned with the idea of literary value. Themes here include the roles of content and form in our appreciation of the work, the relation of character to plot, the possibility of ethical evaluation of literary works, and indeed the very possibility of a coherent notion of literature, thought of as an evaluative category.

When PLF is pursued in this generalizing way, individual works of narrative art are treated as illustrative of general theses about the kinds to which they belong. But this requires a commitment to responsible criticism: criticism which seeks to discover what is in the work, to weigh carefully the evidence—internal and contextual—for different meanings, and to recognize that works of quality rarely fit without residue into categories designed for general explanatory purposes. This may seem an uncontroversial requirement, but advocates of PLF charge that those working in TLF sometimes distort and simplify works and genres in order to fit them into constraining schemas of interpretation such as Lacanian psychoanalysis, while deconstructive scepticism about meaning and its relation to the world has encouraged fantastical readings which convey little insight into the work itself.

So the serious pursuit of PLF requires the philosopher to be also to some extent a critic, following whatever interpretive methods are appropriate to the work in hand. An illustration of the delicate balance between general argument and attention to the specifics of works is given by the debate over narrative unreliability, a topic of concern for students of both literature and film. Narrative unreliability has for long been understood as the product of an unreliable narrator, and examples of this kind are certainly prominent in literature, as with *The Turn of the Screw, The Good Soldier,* and *Pale Fire*. Filmic narratives have also been said to be unreliable, not merely in the sense that they temporarily mislead us as to the situation depicted—as with *The Others* or *Fight Club*—but in the sense that even an attentive viewer may come away from the film with a radically mistaken impression of what has happened in the story, as George Wilson has argued concerning Fritz Lang's *You Only Live Once*.[14] Does this mean that we ought to acknowledge the place in film of a narrator, thought of as an agent who is presenting us with sights and sounds in a way that parallels the role of the governess in presenting us with her account in *Turn of the Screw*? The idea of such a filmic narrator has long been represented in the thinking of practitioners of TLF as the idea of *le Grand Imagier* due to Christian Metz, and in PLF George Wilson advocates a version of this theory.[15] The debate over whether filmic unreliability requires a filmic narrator and, more broadly, whether there are reasons to postulate such narrators as a universal or at least standard component in the narrative package for both literature and film is currently in dispute in ways that draw heavily on the interpretation of particular works. For a primary constraint on theories of narration is that they should account satisfactorily for unorthodox or transgressive narratives as well as for conventional ones.[16]

When philosophers of literature and film consider particular works they aim, as I have indicated, to do justice to their individuality; the result is sometimes a detailed interpretation of the work in question not clearly distinguishable from what might be done by a

[14] See Wilson (1986). See also Wilson (2011, part III). On *The Others, Fight Club,* and other "epistemically twisted" movies, see Wilson (2011, ch. 7).

[15] See Wilson (2011, part II).

[16] See Currie (2004, ch.7) and Kania (2005).

literary or filmic scholar whose primary interest is in enriching our understanding of the dramatic tropes, themes, affective resources, and characterizations of a work.[17] The question arises, then, as to whether this sort of "deep" interpretation is better left to literary scholars, on whose work the philosophers may draw. No doubt philosophers need to be careful here, and must be prepared to learn and respect the ways of scholarship; they also draw freely on the interpretive work of more specialized scholars to illustrate and test their more general claims. But there is value in the treatment of certain narrative works by philosophers; their skills and talents are often suited to the careful interrogation of texts, and the distinguishing of spurious associations and symbolic posturing from insights of real worth. Sometimes it is the philosopher who has the critical and descriptive vocabulary required to encapsulate the work's perspective in illuminating terms; we shall see an example of this in the next section.

5. Literature and Film as Philosophy

The question of whether and in what ways philosophers ought to focus on particular works is given extra urgency by the claim that literary and filmic works may themselves be sources of philosophical ideas and, sometimes, contributions to philosophy itself; in that case philosophers may see it as their business to interpret and assess that philosophical content. Is the claim a true one?

The idea that literature may interestingly convey what are in some broad sense philosophical ideas is virtually a truism if we recognize the dialogues of Plato, Berkeley, and Hume as literature. The dialogue form as used in these projects is undoubtedly in the realms of both literature and fiction (we are not asked to believe that Socrates said exactly what the dialogue reports). But it needs to be treated as a special case on account of its capacity simply to represent conventional argument; quite different considerations would be needed to show that the novels of Henry James and the films of Ingmar Bergman have philosophical content.

Leaving aside the dialogue form, the view that literature is able to achieve philosophical results comes in various versions. A weak version says merely that works in these media may *provoke* philosophical thought and argument, as reading or watching *The Third Man* might generate a discussion on the moral limits of loyalty to friends.[18] Virtually any story with human or human-like characters has this capacity, and so may be counted as philosophical in this weak sense. But works may count as more than trivially philosophical by supporting and perhaps *guiding* philosophical thought in interesting ways through their narrative structures, delineation of characters, and use of point of view; for many, the later work of Henry James provides the supreme instances of this sort of philosophical literature.

Some works of literature offer explicit philosophical commentary, as with Proust on time and memory; voice-over gives film the same capacity, though it is not often exploited to that purpose. But the distinction between explicit and implicit philosophical content is

[17] See e.g. Isenberg (1951), Currie (2009), Wilson (1986).

[18] Carol Reed's film appeared in 1949 and Graham Green's novella from his own script in 1950.

less crucial, methodologically, than it might seem. As we have learned from pragmatics, explicit content, even of a quite mundane kind, is rarely *wholly* explicit, and filling it out is an interpretive enterprise; the communicative import of "he has money" will be understood quite differently in response to a question about (a) who will pay for the bus fares, and (b) why Mary would consider marrying that unappealing person.[19] And however explicit the statement is, we need to make decisions about whether it, or some part of it, is meant ironically, or is in the mouth of an unreliable narrator. The existence of an explicit and apparently philosophical remark is never more than evidence, and sometimes rather weak evidence, for the work's assertion of a philosophical view, while implicit content is sometimes rather easy to arrive at, given the rhetorical and other aspects of the fictional narrative. The distinction between what is explicit and what is implicit—to the extent that this is a viable distinction at all—is not the distinction between communicative content that is manifest and content which requires interpretation.

A stronger claim still about the philosophical import of film and literature is that some works have a capacity to embody philosophical ideas in ways which are both important and unmatched by other sources: on this view no other activity—discursive argument for example—can fulfil the role of philosophical literature and film. This has been argued for, in the case of literature, by Martha Nussbaum who claims that narrative fictions of quality are uniquely able to help us to recognize and explore the complexities of moral choice.[20] Those who take this approach to the philosophy of literature and film may defend it in either of two ways. One is to provide general arguments for the claim that high-quality works are especially suited to convey philosophical content; the other is to show the particular philosophical content that particular works have. Frequently, these projects are pursued together, as they are in Nussbaum's *Love's Knowledge*. In other cases, the greater emphasis is on establishing the richness of the particular work's insights, as with Mulhall's reflections on the four *Alien* films.[21]

While the claim that literature and film make an essential contribution to philosophy is difficult to establish, or perhaps even to rationally assess, a good deal of work has borne witness to the philosophical interest that certain narratives do have. And in the process of drawing out this philosophical meaning interpreters have helped us to understand specific and idiosyncratic aspects of the works themselves; here the generalizing and particularizing tendencies in PLF are hardest to separate. Thus Paisley Livingston undertakes to show that some of Bergman's films reflect the influence of a specific philosophical position—that of the Swedish philosopher Kalia; in the process we are exposed to novel ways of understanding such films as *Persona*.[22] Exemplary in this genre is the commentary provoked from philosophers by the film *Memento*; the film's themes of memory and responsibility would be particularly hard to disentangle without analytical skill and training. In an elegant essay Joseph Levine elaborates a view of the architecture of memory which helps to explain the main character's defective epistemic position; he also finds in these considerations grounds for rejecting the thesis of the extended mind.[23] The philosophy here does not

[19] Relevance Theorists have made a major contribution to our understanding of the ways in which pragmatics plays a role in understanding what is said in an utterance as well as in determining what is implicated by the utterance. See e.g. Sperber and Wilson (1986), and Carston (2002).

[20] Nussbaum (1990). [21] Mulhall (2002). [22] Livingston (2009).

[23] Levine (2009).

seem like an optional extra, to be enjoyed separately from one's experience of the film. The essay offers a genuine interpretation of the story—a way of understanding the predicament of the film's central character which is perhaps not evident in the filmic structure itself, but to which the work is hospitable. In that case we have an example of a distinctly philosophical interpretation of a work.

6. Darwinian Approaches

Our conclusions so far are that PLF combines traditional, more or less a priori philosophical argument with the interpretive methods appropriate to close textual analysis. A more radical suggestion is that PLF should also embrace the systematically empirical methods of the sciences. The rest of this chapter will be devoted to assessing this idea. One reason for considering it is that there has recently been work on the question whether our interest (apparently universal) in stories is explicable in terms of the forces of natural and sexual selection which shaped our species, inviting the thought that a comprehensive theory of literature and film will be in part Darwinian. Practitioners of PLF keen to advance this program must become skilled in deploying arguments about inclusive fitness, evolutionary game theory, honest signalling, and (perhaps most relevantly) gene-culture co-evolution.

This immediately presents a difficulty: once we get beyond the basic picture of inheritance of differentially distributed traits with different contributions to fitness, the current neo-Darwinian picture of evolution is alarmingly complex and hence difficult to embed within works aimed at (or written by) a philosophical community with little grounding in these issues. Another problem is that, while we have quite sophisticated theoretical tools for understanding Darwinian selection we know much less about the evolution of our species than we would like; this leaves speculation about the supposed evolutionary origins of this or that cultural practice worryingly unconstrained.

So far, work in this area may be described as suggestive at best and sometimes prone to over-reach itself, especially when the aim is to shed light on the qualities of particular works. Brian Boyd presents a wide-ranging thesis about the evolutionary advantages of the arts generally, but with special attention paid to fictional narrative; he argues that the evolutionary function of fiction is to improve our processing of social information—a topic I return to in the final section.[24] In a lengthy analysis of themes from the *Odyssey*, he suggests that an evolutionary approach can shed light on the story's structure and emphasis: "an awareness of the problems of the evolution of cooperation explains why [the Odyssey] ends where it does", that is, with the extensive post-reunion material which some have found puzzling.[25] But questions about the appropriate balance of material within a plot seem to be resistant to this kind of treatment: why should someone who thinks the plot unbalanced be persuaded out of that view by arguments about the forces at work in our evolutionary past?

[24] Boyd (2009). For a somewhat critical review, especially regarding the lack of empirical evidence (for which see the final section of this chapter), see Verpooten (2011).

[25] Boyd (2009), p. 317.

Some work in this genre aims to shift our ways of categorizing narrative works by revealing themes and commonalities which would otherwise be hidden or seem insignificant. William Flesch focuses on the relation between plot and evolutionarily important problems faced by our ancestors, in particular the kind of altruistic or costly punishment he says we often find in literature, where a deceiver is exposed by an altruistic third party whose efforts and sacrifices can't be explained in selfish terms. Such costly acts are, says Flesch, crucial to binding communities together, and their performance is compensated for by the admiration they provoke in the rest of us, securing the punisher a favourable reputation and opportunities for co-operation (thus making the punishment not costly after all). This admiration is so natural to us that it is provoked by purely imaginary scenarios of costly punishment—hence literature.[26]

One problem here is exactly how much literature actually conforms to the costly punishment plot; another is the extent to which costly punishment figures as a significant force in well-developed evolutionary theories. But perhaps Flesch is simply focusing on a too limited and atypical range of cognitive/emotional responses which fiction can unleash. One might say, rather more generally, that literature grants us imagined access to that which in reality is difficult to come by: the motives of others. Mindreading is a capacity we have been granted, in modest degree, by Darwinian natural and sexual selection as a response to the problems of co-operating and competing with others in intensely social environments. But in the wild, mind-reading is hard work, gets limited results, and may fail altogether because our target is cannily deceptive about her intentions. Responding to this thought, one might argue that much of what goes on in literature and in linguistically mediated fiction generally can be understood as providing us with a fantasy of insight: a world in which we are able with ease to monitor the thoughts, and hence the plottings and defections, of others. Just as our highly focused attention to sexual characteristics makes erotica and pornography possible, so our attunement to motive makes for the literature of love, rivalry, and misunderstanding, which is virtually all the literature we have. While film and television provide fewer opportunities for the direct display of inner states, their ability to concentrate on visual detail makes them tempting vehicles for another kind of fantasy-of-understanding: the easy reading of expression. As Lisa Zunshine has pointed out, facial expressions are often very ambiguous in reality and are easily misunderstood; visual media of narration often provide us with characters whose expressions are pleasingly but unrealistically revelatory of their inner states.[27]

Those who use arguments such as these to advocate Darwinian methods in PLF will need to answer the following question: assuming that facts about our desire for insight into the minds of others and the difficulty of satisfying those desires are relevant to the understanding and appreciation of literature and film, why should this be considered an argument for a Darwinian turn in PLF? Surely what matters is that we have these desires and that they are, as a matter of fact, hard to satisfy; whether we got to this position by Darwinian selection or by the whim of a creator is irrelevant.[28] One answer to the objection agrees that, so far as the

[26] See Flesch (2007).

[27] See Zunshine (2012). Other works on the relation between literature and evolutionary theory include Zunshine (2006), Gottschall (2008), J. Carroll (1995), (2004), (2011), Boyd (2009), Vermule (2010), Easterlin (2012), Gottschall and Sloan Wilson, (2005).

[28] See Currie (2004, ch.13).

philosophy of criticism is concerned, we need not concern ourselves with the natural history of our tendencies to be engaged, delighted, or disappointed by works of narrative art; for the philosophy of criticism focuses on the ways in which aesthetic judgements about works and classes of works can be justified, and we do not justify any aesthetic response by pointing to its natural history. But the philosophy of literature and film is (or ought to be) a much larger enterprise which seeks to give the best overall picture of the phenomena it studies, and part of that best overall theory will be an account of the evolutionary basis (if there is one) of our attraction to these forms of narration. This of course does not mean that every philosopher of literature and film need be concerned with this aspect of the overall project.

7. Claims to Knowledge

The Darwinian approach seeks to put new questions on the agenda of PLF; to the extent that those questions are resisted, we may feel no obligation to abandon the traditional approach: a mixture of a priori philosophical argument and sensitivity to the meanings and values of particular works. In this final section I suggest that there is a question at the heart of the project of PLF—a question familiar to us from Plato and Aristotle on—which can only sensibly be answered by close attention to empirical results in the sciences of mind, but which has, by and large, been treated as if it were amenable to those traditional methods.

One claim that is often heard concerning the value of literature is that it is a source of knowledge; similar claims are sometimes made about films. This is, once again, a substantive rather than a methodological question, but in considering how to answer it we raise important questions about method, and in particular about the relations between philosophy and empirical inquiry.

We have already touched on the question of literature's contribution to philosophy, a contribution which, if real, should in some way promote the growth of knowledge. Here I examine a claim about literature's capacity to provide insight of a kind which is part philosophical, part psychological, and part practical: the nature and workings of the mind and its relations to the limits and complexities of moral decision and action. Shakespearean drama and the great novels of the nineteenth century are often said to be paradigmatic of this capacity to illuminate the otherwise obscure realities of our ethical relations to ourselves and to others.[29]

This is certainly in part a traditionally philosophical doctrine. For example, there is a good deal of conceptual work to be done in distinguishing between relevant kinds of knowledge. It has been suggested that the cognitive value of literature lies not so much in the truths it teaches as in the skills it promotes, making us more interpersonally insightful and sensitive than we would otherwise be.[30] We need, therefore, a clear understanding of the relations between knowing-that and knowing-how. But whatever version one adopts

[29] See e.g. Carroll (1998), Currie (1995, 1998), Diamond (1983), Gaut (2007), Gibson (2012), Goldberg (1993), Goldman (2013), Jacobson (1996), Nussbaum (1986), (1990), Palmer (1992), Pippin (2001).

[30] For a sceptical view of the novel's capacity to make us more empathic, see Keen (2007).

of the claim that literature is improving, this debate cannot be conducted in wholly a priori terms: we need to discover how and under what circumstances literature enhances our knowledge (of whatever kind), whether it sometimes (perhaps more often) leads to error and ignorance, and what kinds of people most easily benefit from what kinds of literature. These are empirical questions, but the approach that philosophers have taken to them contrasts starkly with the approach to a closely related topic: the negative effects of fiction. Those who condemn various kinds of fiction as leading to bad behaviour are rightly challenged to produce evidence for their claims, and the few philosophers who have taken up this cause have generally been careful to find evidential support for their case.[31] It is not uncommon, by contrast, for someone to write a long book extolling the civilizing powers of fiction, and to fail completely to provide, or even consider providing, any evidence for the claim.

Part of the explanation may be that there is an implicit appeal in these positive arguments to the evidence supposedly manifest in the act of making those arguments; the author manages to suggest that her own manifestly civilized values and demeanour are evidence enough for the claim, whereas one would not readily put oneself forward as an instance of the power of fiction to degrade and corrupt. We should not automatically discount people's claims to know that literature has changed them or to recognize such change in others, but there are reasons to be careful about such claims. There are general reasons for doubting that people have reliable access to their own inner processes of learning; we also know that people are very apt to believe in the positive worth of things they have chosen to invest in.[32] It is quite possible, then, that, rather than reporting realistically on their own development, the advocates of literature are persuaded by their own self-supporting but false account. And if the work on literature's moral and epistemic value is read mostly by those with a commitment to reading, we can expect these accounts to receive a relatively uncritical welcome.

These observations suggest that practitioners of PLF would do well to look for more systematic experimental evidence for their claims. Unfortunately, the little evidence we currently have is somewhat ambiguous. While the effects of smoking on health are more or less established, we have not found ways to untangle the civilizing effects (if any) of literature from those of, say, an otherwise advantaged but unliterary upbringing. It is not easy to figure out either the direction of causation, since being, say, highly empathic might be a cause of reading literature rather than an effect, or to rule out a common cause of high levels of both empathy and literary engagement. The evidence that exists seems to be at its most reliable in dealing with questions about factual knowledge. Thus there is good evidence that readers are apt to "import" propositions about history, geography, and other matters from fictions, either when such information forms the implicit background to the story, or when a character presents the information in a conversation and is not contradicted. Often, however, the imported information is false: subjects' credences for propositions such as "Chocolate makes you slimmer" rise when they are presented with a short story in which a

[31] See Hurley (2006).

[32] On the failure of introspection, see Carruthers (2011). In a study about which I say more later, Kahneman and Klein (2009)(the latter of whom has a generally pro-attitude to expert judgement) agree that one thing that is not predictive of expert competence is the subject's own confidence in their judgement.

character makes this implausible claim.[33] There is even some evidence that people are more willing to absorb this kind of information from avowedly fictional texts than they are from texts announced as non-fictional.[34] The effects of fiction on belief are particularly well documented for cases of sustained viewing of TV serials: people have picked up a great deal of medical information from *ER*, for instance.[35]

A small amount of work has also been done that speaks to the question of change in evaluative opinion wrought by fiction: having read a brief fictional story of a particularly downbeat kind, readers seem to be more hospitable than they previously were to such ideas as "The world is a dangerous place".[36] Whether this represents genuine change of opinion rather than, say, mood is unclear given the lack of evidence that such changes last beyond days or even hours.[37] Taken as a whole, the evidence certainly supports the proposition that literature and film (and fiction more generally) are capable of having various effects on the beliefs and other mental states of audiences. It is much more difficult to find strong evidence for a correlation between the quality of the works concerned and the educative, civilizing, or otherwise positive character of their effects.[38]

The fact that evidence is currently wanting is of course no defence against the charge that one is insensitive to the need for evidence. But one might defend the philosophers of art here by saying that these two projects—the philosophical and the psychological—seek to answer different questions, and their different methods and conclusions are therefore of no concern. The philosophical project is concerned with elucidating the capacity of the best narrative art to educate us, if we are prepared, by dint of knowledge and attention, to meet its challenge. The psychological project, on the other hand, is concerned simply to measure the effect of brief and usually artificially devised fictions on statistically informative numbers of uncommitted and not well prepared subjects.

Two things need to be said in response to this. The first is that one would be in an uncomfortable position in restricting claims about the value of literature to an especially atypical and hard to test group of subjects, while arguing—as advocates of literature often do—that literature has a civilizing role in the school curriculum. The second is that, however restricted, claims about the capacity of something to bring about cognitive and behavioral change in somebody need to be verified by observation of the relevant cognitive or behavioural change. Philosophers are no better able to undertake those observations for this special group than they are for the population as a whole.

[33] Gerrig and Prentice (1991). For philosophical commentary on some of this research, see Friend, forthcoming.

[34] Prentice and Bailis (1999); for a contrary view, see Schrum et al. (1998) and Richert and Smith (2011).

[35] See Brodie et al. (2001).

[36] See Green and Brock (2000).

[37] For a review of studies up to 1999, see Hakemulder (2000).

[38] A very recent study which has gained some attention for its claim that reading fictions of quality contributes to competence in "theory of mind" is Kidd and Castano 2013. For criticism of its methodology, see the comments of linguist Mark Liberman at <http://languagelog.ldc.upenn.edu/nll/?p=7715> (accessed September 26, 2015). Whatever the ultimate value of this study, this is surely the kind of work that philosophers interested in the relation between fictional narratives of quality and knowledge should attend to. See also Marr et al. 2011. For a more cautious approach, see Koopman 2015.

This second response might be countered in the following way. The philosopher's claim is that there are things we can learn from literature, and the philosophical project is completed when we show, by close analysis of the relevant works, what it is that can be learned from them. It is this interpretive project which provides the relevant evidence, and we do not need to look to the results of psychological experiments. This response fails for two reasons. The first is that no one is claiming that a philosophical elucidation of the literary work's message can make manifest to us its educative capacity, for no one is claiming that what the fiction provides is merely some set of propositions we might come to know about from a moral tract or psychological text. For example, Nussbaum's claim is that it is the narrative *form* which makes works of literature especially valuable in the cognitive and moral domains, and a claim about the efficacy of form cannot be vindicated by the recitation of content. The second is that propositional knowledge, from whatever source, is not all that is at issue in the debate over learning from literature; as I have indicated, it is also generally agreed to be partly a matter of skill or know-how. It is always a difficult, and certainly an empirical, question as to how skills are best acquired or improved, and the claim that fictions in general are, or some particular fiction is, well suited to this task can be vindicated only by showing that people learn the skill better from this source than from other sources. Again, the recitation of a work's content (assuming that such a thing is possible) does nothing to show this.

None of this is meant to suggest that philosophers should stop theorizing about the moral and epistemic value of literature. There is work to be done in developing theories of how we *might* learn from fiction, especially given that psychological work in this area sometimes suffers from an impoverished view of the explanatory options. But philosophers should be guarded in their claims as long as they have no direct evidence in their favour, and while there exists a considerable body of indirect evidence which speaks against them. They should have a lively and up-to-date awareness of what the state of the evidence is, developing theories which have the best chance of surviving crucial tests, and they should try to frame their hypotheses in ways that suggest how these tests might be carried out. At the very least they ought to recognize that any claim worth attending to in this area must be empirically testable.

Bibliography

Bordwell, D. and Carroll, N. (1996) *Post-Theory: Reconstructing Film Studies*. Madison, University of Wisconsin Press.

Boyd, B. (2009) *On the Origins of Stories: Evolution, Cognition and Fiction.* Cambridge, MA: Belknap Press of Harvard University Press.

Brodie, M., Foehr, U., Rideout, V., Baer, N., Miller, C., Flournoy, R., and Altman,D. (2001) Communicating Health Information Through The Entertainment Media. *Health Affairs* 20(1): 192–9.

Carroll, Joseph. (1995) *Evolution and Literary Theory.* Columbia, MO, University of Missouri Press.

Carroll, Joseph. (2004) *Literary Darwinism: Evolution, Human Nature, and Literature.* London and New York, Routledge.

Carroll, Joseph. (2011) *Reading Human Nature: Literary Darwinism in Theory and Practice*. New York, SUNY Press.
Carroll, N. (1988) *Philosophical Problems of Classical Film Theory*. Princeton, NJ: Princeton University Press.
Carroll, N. (1997) Fiction, non-fiction and the film of presumptive assertion. In: R. Allen and M. Smith (eds.), *Film Theory and Philosophy*, Oxford, Clarendon Press, 173–202.
Carroll, N. (1998) Art, Narrative and moral understanding. In: J. Levinson (ed.), *Aesthetics and Ethics: Essays at the Intersection*, Cambridge, Cambridge University Press, 126–60.
Carruthers, P. (2011) *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford, Oxford University Press.
Carston, R. (2002) *Thoughts and Utterances: The Pragmatics of Explicit Communication*. Oxford, Blackwell.
Cavell, S. (1971) *The World Viewed: Reflections on the Ontology of Film*. New York, The Viking Press; 2nd enlarged edition (1979), Cambridge MA, Harvard University Press.
Cavell, S. (1979) *The Claim of Reason: Wittgenstein, Skepticism, Morality and Tragedy*. New York: Oxford University Press.
Cavell, S. (1981) *Pursuits of Happiness: The Hollywood Comedy of Remarriage*. Cambridge, MA, Harvard University Press.
Cavell, S. (1987) *Disowning Knowledge: In Six Plays of Shakespeare*, Cambridge: Cambridge University Press; 2nd enlarged edition (2003): *Disowning Knowledge: In Seven Plays of Shakespeare*.
Currie, G. (1995) *Image and Mind: Film, Philosophy, and Cognitive Science*. Cambridge, Cambridge University Press.
Currie, G. (1999) Visible Traces: Documentary and the Contents of Photographs. *Journal of Aesthetics and Art Criticism* 57(3): 285–97.
Currie, G. (2004) *Arts and Minds*. Oxford, Clarendon Press.
Currie, G. (2009) Agency and repentance in *The Winter's Tale*. In: M. D. Bristol (ed.), *Shakespeare and Moral Agency*. London, Continuum.
Currie, G. (2010) *Narratives and Narrators*. Oxford, Oxford University Press.
Currie, G. (2012) Literature and truthfulness. In: J. Maclaurin (ed.), *Rationis Defensor: Essays in Honour of Colin Cheyne*. Dordrecht, Springer, 23–31.
Currie, G. and Jureidini, J. (2004) Narrative and Coherence. *Mind & Language* 19(4):409–27.
Devitt, M. and Sterelny, K. (1999) *Language and Reality: An Introduction to the Philosophy of Language*. 2nd ed. Oxford, Blackwell.
Diamond, C. (1983) Having a Rough Story about What Moral Philosophy Is. *New Literary History* 15: 155–69.
Easterlin, Nancy. 2012. *A Biocultural Approach to Literary Theory and Interpretation*. Baltimore, Hopkins University Press.
Flesch, W. (2007) *Comeuppance: Costly Signalling, Altruistic Punishment, and Other Biological Components of Fiction*. Cambridge, MA, Harvard University Press.
Friend, S. (2012) Fiction as a Genre, *Proceedings of the Aristotelian Society* 112:179–209.
Friend, S. (forthcoming) Believing in stories. In G. Currie, M. Kieran, A. Meskin and J. Robson (eds), *Aesthetics and the Challenge from the Sciences*, Oxford, Oxford University Press.
Gaut, B. (2007) *Art, Emotion and Ethics*. Oxford, Clarendon Press.
Gerrig, R. J. and Prentice, D. B. (1991) The Representation of Fictional Information. *Psychological Science* 2(5): 336–40.

Gibson, J. (2012) Fiction and the Weave of Life. Oxford, Oxford University Press.
Goldberg, S. L. (1993) Agents and Lives: Moral Thinking in Literature. Cambridge, Cambridge University Press.
Goldman, A. (2013) *Philosophy and the Novel*. Oxford, Oxford University Press.
Gottschall, J. (2008) *The Rape of Troy: Evolution, Violence and the World of Homer*. New York, Cambridge University Press.
Gottschall, Jonathan and Wilson, David Sloan. (eds.) (2005) *The Literary Animal: Evolution and the Nature of Narrative*. Evanston, IL, Northwestern University Press.
Green, M. C. and Brock, T. C. (2000) The Role of Transportation in the Persuasiveness of Public Narratives. *Journal of Personality and Social Psychology* 79: 701–21.
Grove, W. M., and Lloyd, M. (2006) Meehl's Contribution to Clinical Versus Statistical Prediction. *Journal of Abnormal Psychology* 115 (2): 192.
Hakemulder, J. (2000) *The Moral Laboratory: Experiments Examining the Effects of Reading Literature on Social Perception and Moral Self-Concept*. Philadelphia: Amersterdam, John Benjamins.
Harman, G. (1979) Eco-location. In: G. Mast and M. Cohen (eds.), *Film Theory and Criticism*. 2nd ed. New York, Oxford University Press. (This essay did not appear in subsequent editions.)
Hume, D. (1777/1985) On tragedy. In: E. F. Miller (ed.), *Essays Moral, Political and Literary*. Indianapolis: Liberty fund, 216–25.
Hurley, S. (2006) Bypassing conscious control: Media violence, unconscious imitation, and freedom of speech. In: S. Pockett, W. Banks, and S. Gallagher (eds.), *Does Consciousness Cause Behavior? An Investigation of the Nature of Volition*, Cambridge, MA, MIT Press, 301–38.
Isenberg, A. (1951) Cordelia Absent. *Shakespeare Quarterly* 2(3):185–94.
Jacobson, D. (1996) Sir Philip Sidney's Dilemma: On the Ethical Function of Narrative Art. *The Journal of Aesthetics and Art Criticism* 54(4):327–36.
Johnson, S. (1765) *Preface to the Plays of William Shakespeare*.
Kahneman, D. and Klein, G. (2009) Conditions for Intuitive Expertise: A Failure to Disagree. *American Psychologist* 64(6):515–26.
Kania, A. (2005) Against the Ubiquity of Fictional Narrators. *The Journal of Aesthetics and Art Criticism* 63(1): 47–54.
Keen, S. (2007) *Empathy and the Novel*. New York, Oxford University Press.
Kidd, D. and Castano, E. (2013) Reading Literary Fiction Improves Theory of Mind. *Science* 342(6156):377–80.
Kivy, P. (1997) *Philosophies of Arts: An Essay in Difference*. New York, Cambridge University Press.
Koopman, E. (2015) Empathetic Reactions after Reading: The Role of Genre, Personal Factors and Affective Responses. *Poetics* 50: 62–79.
Lamarque, P. (2008) *The Philosophy of Literature*. Oxford, Wiley-Blackwell.
Lamarque, P. and Olsen, S. H. (1994) *Truth, Fiction and Literature: A Philosophical Perspective*. Oxford: Clarendon Press.
Levine, J. (2009) Leonard's system: Why doesn't it work? In: A.Kania (ed.), *Memento*, Oxford: Routledge.
Livingston, P. (2009) *Cinema, Philosophy, Bergman: On Film as Philosophy*. Oxford, Oxford University Press.

Mar, R., Oatley, K., Djikic, M., and Mullin, J. (2011). Emotion and Narrative Fiction: Interactive Influences Before, During, and After Reading. *Cognition and Emotion* 25: 818–33.

Meehl, P., (1954/1996) *Clinical vs. Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. Northvale, NJ, Jason Aronson Inc.

Meskin, A. (2009) Comics as Literature? *British Journal of Aesthetics* 49: 219–39.

Mulhall, S. (2002) *On Film*. 2nd ed. London and New York, Routledge.

Nussbaum, M. (1986) *The Fragility of Goodness: Luck and Ethics in Greek Tragedy on Philosophy*. Cambridge, Cambridge University Press, 1986.

Nussbaum, M. C. (1990) *Love's Knowledge: Essays and Philosophy on Literature*. Oxford, Oxford University Press.

Palmer, F. (1992) *Literature and Moral Understanding*. Oxford, Oxford University Press.

Pippin, R. (2001) *Henry James and Modern Moral Life*. Cambridge, Cambridge University Press.

Prentice, D. A. and Gerrig, R. J. (1999) Exploring the boundary between fiction and reality. In: S. Chaiken and Y. Trope (eds.), *Dual-Processing Theories in Social Psychology*, New York, Guildford Press, 529–46.

Richert, R. A. and Smith, E. I. (2011) Preschoolers' Quarantining of Fantasy Stories. *Child Development* 82(4): 1106–19.

Schrum, L. J., Wier, R. S. and O'Guinn, T. C. (1998) The Effects of Television Consumption on Social Perceptions: The Use of Priming Procedures to Investigate Psychological Processes. *Journal of Consumer Research* 24: 447–58.

Sperber, D. and Wilson, D. (1986) *Relevance: Communication and Cognition*. Oxford, Blackwell.

Trilling, L. (1950) *The Liberal Imagination: Essays on Literature and Society*. New York, The Viking Press.

Vermeule, Blakey. 2010. *Why Do We Care about Literary Characters?* Johns Hopkins University Press.

Verpooten, J. (2011) Brian Boyd's Evolutionary Account of Art: Fiction or Future? *Biological Theory* 6: 176–83.

Walton, K. L. (1984) Transparent Pictures: On the Nature of Photographic Realism. *Critical Inquiry* 11(2): 246–77.

Walton, K. L. (1997) On pictures and photographs: Objections answered. In: R. Allen and M. Smith (eds.), *Film Theory and Philosophy*. Oxford, Oxford University Press, 60–75.

Wilson, G. M. (1986) *Narration in Light: Studies in Cinematic Point of View*. Baltimore, Johns Hopkins University Press.

Wilson, G. M. (2011) *Seeing Fictions in Films: The Epistemology of Movies*. Oxford, Oxford University Press.

Zunshine, L. (2006) *Why We Read Fiction: Theory of Mind and the Novel*. Columbus, Ohio State University Press.

Zunshine, L. (2012) *Getting Inside Your Head*. Baltimore, Johns Hopkins University Press.

CHAPTER 33

AESTHETICS AND PHILOSOPHY OF ART

DOMINIC MCIVER LOPES

1. Introduction

Like the philosophy of science, aesthetics and the philosophy of art treats problems that are typically at home in 'core' areas of philosophy (metaphysics and epistemology, philosophy of language and mind, and value theory), as they apply to specific domains of human activity. Since what is special about these domains constrains solutions to the problems, and even colours our thinking about what the problems are, most if not all work in aesthetics and the philosophy of art reacts to, answers to, or attempts to explain extra-philosophical facts about aesthetic and artistic phenomena. The field is therefore a case study in how philosophy may relate as a neighbour to empirical studies of culture.

2. A Picture of the Field

This chapter deals with aesthetics and the philosophy of art as branches of so-called analytic philosophy. With few exceptions (notably Danto 1986, 1997) and for better or worse, the field has seen little impact from nineteenth-century German philosophy and the major continental figures of the last century, partly because of profound methodological differences (see Schaeffer 2000). Within analytic philosophy, aesthetics and the philosophy of art intersect and they are nowadays pursued in tandem, but they are distinct. Aesthetics is concerned, at bottom, with a distinctive type of value as it is realized by a wide range of items, including natural environments, designed artefacts and events, the elements of everyday life, and non-artistic representations (e.g. scientific theories), as well as works of art. Philosophy of art, which spun off aesthetics in the 1960s, concerns the nature of art and the individual arts, the ontological categories to which art works belong, the mechanisms art works employ (e.g. depiction and fiction), the various kinds of cognition that art works

demand of their makers and consumers (e.g. creativity and interpretation), and the values that are realized by works of art, including but not limited to aesthetic value. This division of the field has a historical dimension, and it shapes philosophical interactions with other disciplines.

Many of the topics making up aesthetics obviously date back centuries. Plato proposed a theory of beauty, or aesthetic value, principally realized by theories and human bodies, though he took a famously dim view of the value of what we would now call the representational arts. Aristotle, responding in defence of theatre, may be credited with the earliest account of a specific art form. The medievals kept up the ancient interest in beauty by connecting it to theological concerns, and they explored human creativity as an analogue of a divine creativity. Beauty and taste, the faculty for apprehending beauty, topped the agenda of early modern aesthetics, notably in the writings of Hutcheson, Shaftesbury, Hume, Burke, Reid, Baumgarten, and Kant.

The philosophy of art does not appear to have nearly the same vintage; Aristotle's *Poetics* and medieval writing on creativity stand out as exceptions to the long dominance of the aesthetic. Meticulous research by the historian Paul Oskar Kristeller (1951–2) traces the conceptual origins of the philosophy of art (see also Shiner 2001). Kristeller shows that it was not until the mid-eighteenth century that a concept of art came to group some activities, namely the fine arts, together with each other and apart from such other activities as the liberal arts, sciences, applied sciences, and crafts. Thus ancient Greek and Latin uses of '*technē*' and '*ars*' encompass a range of activities that we would now classify not as arts but rather as skills, crafts, and sciences (Kristeller 1951–2, 498–9). According to the influential scheme proposed by Hugo of St Victor in the twelfth century, architecture, sculpture, and painting are grouped together under *armatura*, music is a branch of mathematics, and poetry belongs with grammar, rhetoric, and logic (Kristeller 1951–2, 507–8). Chambers's *Cyclopedia* of 1727 classifies painting with optics, music under applied mathematics, gardening under agriculture, and poetry alongside rhetoric, grammar, and heraldry (Kristeller 1951–2, 520). Only a few decades later, Diderot's *Encyclopédie* groups the fine arts as is now familiar. The change is dramatically graphed in tree diagrams that accompanied the two encyclopaedias. Of course, the concept of art did not appear overnight: it developed over many decades, as a result of a number of historical factors, which Kristeller identifies (1951–2, 510–26). However, it was finally introduced as a theoretical concept in the salons and seminar rooms of Paris, whence it quickly made its way into the repertoire of folk concepts.

In so far as the philosophy of art bears upon a sphere of human culture, it cannot have antedated the conception of that sphere. Indeed, the sphere was only seen to come together through the introduction of a philosophical theory of the arts (Batteux 1746). This event had momentous consequences for the disciplinary organization of the modern university in the nineteenth century. The fine arts are studied and taught in schools of music, architecture, and landscape architecture and in departments of literature, theatre, film, dance, and art history. There are no academic disciplines devoted to fashion, interior decoration, jokes, typography, or vexillology.

A consequence of this history is that the philosophy of art shares a first-order subject matter with musicology and music theory, linguistics and literary theory, architectural history, the anthropology of dance, the sociology of visual art, and other art-oriented

scholarship. Meanwhile, philosophical work in aesthetics shares a first-order subject matter with social and behavioural studies of aesthetic response and with a small number of disciplines, such as oenology, which treat artefacts that have not achieved art status.

A side effect of this history is that psychologists and neuroscientists study what are really aesthetic phenomena though they often tend to focus on art works. For example, a recent spate of books about visual art written by vision scientists actually explain responses to visual design that are not limited to art (e.g. Zeki 2000, Livingstone 2002). Arguably the focus on art is less a matter of what methodology requires than marketing that trades on the cachet of art. By the same token, it has been a struggle for philosophers to recall our attention to aesthetic phenomena that have lately been overshadowed by art (e.g. Carlson 2000; Korsmeyer 2002; Brady 2003; Saito 2008).

3 Methodology in Three Dimensions

Writing in aesthetics and the philosophy of art has rarely been forthright about method, but the field is no exception to the trend towards greater methodological awareness in philosophy at large. Recent years have seen several fine efforts—some of them discussed below—to advocate methods that are well suited to the field's subject matter (e.g. Zangwill 1995; Davies 2004, ch. 1; Thomasson 2005; Walton 2007; Schellekens 2011; Currie, Kieran, and Meskin 2014). Several factors drive this trend. One is developments in other areas of philosophy, such as experimental philosophy. Then there is always the hope of clearing gridlocked disputes by reconsidering whether the right questions are being addressed using the right tools. Meanwhile, an explosion of scientific research on art and the aesthetic has sparked direct dialogue between philosophers and scientists, which has included some discussion of the role of philosophy in relation to science (e.g. Palmer and Shimamura 2011). Increasing specialization in the philosophy of art has meant that philosophers are expected to know at least one art form well enough to engage with experts outside philosophy, and this naturally prompts reflection about how philosophy may best contribute to or draw from humanistic research. This complex of factors has led to diverse explorations in method, which can be oriented around a set of orthogonal distinctions, between pure and applied philosophy, internal and external perspectives, and first- and second-order subject matters.

Richard Wollheim took aesthetics and the philosophy of art to be instances of applied rather than pure philosophy (1999, xi–xii). Whereas pure philosophy has conceptual analysis as its method, applied philosophy supplements conceptual analysis with 'whatever else can serve its needs', including observation, experiment, common usage, and traditional lore. Accordingly, pure philosophy aims at conceptual or logical truth whereas applied philosophy aims at 'theoretical necessity', which has the same degree of generality as the laws of nature. And while the subject matter of pure philosophy is perfectly general, applied philosophy characterizes this world or some fragment of it. As Wollheim sums it up, applied philosophy blends into science, so that 'it is a matter more of tact than of observance to recognize where one begins and the other ends' (1999, xi).

Methods in the field afford either internal or external perspectives on aesthetic and artistic phenomena. Although these phenomena figure in the appreciative responses of individuals and also in interpersonal appreciative practices, a description of an aesthetic or artistic phenomenon as it is viewed from outside these practices may not match the description that might be given by practitioners themselves. Peter Lamarque, writing about the accessibility, from the present, of art of the distant (palaeolithic) past, contrasts 'the internal perspective … of the participant for whom the phenomena are endowed with meaning' with the 'external perspective … of the detached observer for whom they are but instances of general sociological laws' (2005, 29). Those taking an internal perspective work out how appreciative practices are structured, provide a rationale for the critical interpretations and evaluations associated with them, and articulate the value that is internal to the perspective. By contrast, philosophers taking an external perspective seek the best overall explanation of the patterns of response that they observe, even if this explanation would be rejected by the participants themselves.

Incidentally, work taking the internal perspective is not necessarily conservative; it may have revisionist ambitions. Indeed, strictly speaking, only the internal perspective permits revisionism, as long as what is open to revision is our practices as we understand them from the inside. Although the external perspective is not necessarily debunking, it may be. When it is, it leaves our practices as it finds them.

Added to the distinctions between pure and applied philosophy and internal and external perspectives is a third distinction, between first-order and second-order subject matter. A method has first-order subject matter when it takes as data some aesthetic or artistic phenomena directly as they appear to the philosophical observer. It has second-order subject matter when it takes as data some aesthetic or artistic phenomena as they figure in the understandings and explanations of empirical research in the humanities or the social and behavioural sciences. The two can be combined, of course. Empirical researchers in different fields may not agree on what it is that they are trying to explain, though it may be beneficial for them to integrate their research programs so that they are trying to explain the same phenomenon. Philosophers can facilitate such integration by bringing empirical explananda into alignment with each other and, sometimes, with an independent philosophical conception of what is to be explained (Bergeron and Lopes 2011).

These three pairs of distinctions are fully orthogonal to each other. Most work in aesthetics and philosophy of art is applied philosophy, though some is pure philosophy. Obviously, applied philosophy affords internal or external perspectives on first- or second-order subject matter, though it is less obvious that these are logically independent of one another. A common mistake is to suppose that a method fitted to a second-order subject matter will afford an external perspective and that a method fitted to a first-order subject matter will afford an internal perspective. Debunking research in the humanities and social and behavioural sciences tends to grab the headlines, but empirical research may sometimes help to articulate the meaning that insiders ascribe to aesthetic and artistic phenomena. A recent example is the appeal to simulation theory as a model to understand fiction and imagination (e.g. Feagin 1995). The model is partly intended to be one that can capture insiders' imaginative experiences and encounters with fiction in terms that they understand. Recent work in feminist aesthetics abounds in examples of the use of first-order

methods with debunking repercussions that destabilize the insider's perspective (e.g. Devereaux 1995; Brand 2000; and Eaton 2013).

4. Traditional Methods

Work in analytic philosophy of art and aesthetics since the middle of the twentieth century has used a considerable range of methods, which fall into two families. One is conceptual analysis, and variants upon it that reflect the special subject matter of the field. The other does not have an established label; call it 'critical demonstration'.

The classic example of critical demonstration is Arnold Isenberg's influential 'Critical Communication' (1949). In this paper, Isenberg takes issue with a conception of art criticism as stating reasons that lend weight to verdicts, for he doubts that a critic's description of, for example, a painting's 'wavelike contour' serves 'to inform us of the presence of a quality as banal and obvious as this'. Instead, he proposes that the description 'gives us directions for perceiving'; it 'narrows down the field of possible visual orientations and guides us in the discrimination of details, the organization of parts, the grouping of discrete objects into patterns' (1949, 336). Moreover, Isenberg's paper is itself an instance of a kind of critical communication. He mounts no argument for his proposal apart from giving rich descriptions of actual examples of art criticism, by means of which he paints a picture of criticism as working in the way he proposes. What is so telling in his picture of art criticism is not so much a fact that lends weight to a conclusion as the way that the fact is described, and there is no further test of the adequacy of the description than its evident aptness.

Critical demonstration is not merely phenomenological description. Monroe Beardsley proposed that aesthetic experience has a distinctive phenomenal character that is evident through introspection (e.g. Beardsley 1981). However, unlike Isenberg, he holds his proposal to a standard of extensional adequacy, and it was in response to an unrelenting barrage of counterexamples that, over several decades, he ended up proposing a series of theories of aesthetic experience. Isenberg's claims in 'Critical Communication' are not established in this way, for critical demonstrations are supposed to be sufficient to guide us to the truth of philosophical claims.

More recent examples of work that falls within the Isenbergian tradition are Martha Nussbaum in *Love's Knowledge* (1992) and Alexander Nehamas in *Only a Promise of Happiness* (2007). One theme of the latter is that the denigration of beauty (and reactionary defences of it) in art and art studies since the nineteenth century rests on an overly narrow conception of beauty. Nehamas arrives at an alternative conception of beauty through a thoughtful engagement with specific art works, which supply a vocabulary giving voice to an apt description of beauty. He speaks, for example, of a 'forward-looking element and the risks that attend it [that] are essential to beauty, which withers when it can promise nothing it has not given already, and signals the fading of love' (2007, 62–3). What this means is only fully expressed in the context of what he has to say about his chosen art works, and especially about his enduring fascination with Manet's *Olympia*. Nehamas offers a critical demonstration whose proof lies in the experience it gives us of these works. For Isenberg, Nussbaum, and Nehamas, one way philosophy is done is by doing a kind of art criticism.

The relative rarity of this kind of writing in analytic aesthetics and the philosophy of art is probably explained by the special skills that are required of its practitioners (and indeed it is less rare in continental philosophy). The family of methods that go under the heading of 'conceptual analysis' call upon a more familiar philosophical skill set.

The armchair variant of the method is championed in George Dickie's (1962) attack on the relevance of empirical research on aesthetic preferences to philosophical treatments of aesthetic norms. Dickie wrote that 'the mechanisms involved in the appreciation of art are similar to such concepts as knowing, believing, oughtness—concepts which all mature users of the language know how to use…. In the case of preference or of the nature of aesthetic experience, we already know what we need to know' (1962, 300–1). The philosopher offers theories of concepts that fit judgements about cases (a.k.a. 'intuitions') that manifest competence with the concepts in question. Nothing more than conceptual competence is needed to generate the judgements about a phenomenon that indicate the correct theory of the phenomenon. So practised, aesthetics is as close as it gets to pure philosophy. Not surprisingly, it is hard to find illustrations of this programme in aesthetics or the philosophy of art (see also Walton 2007, 152).

Far more common is a variant of the method that better fits the field's remit as applied philosophy. This variant takes the phenomena to which it is applied to be elements of, and constituted by, social practices. Provided that the philosopher is a competent participant in the practice, she may devise theories of a phenomenon that fit her judgements about cases, qua participant in the practice (see Thomasson 2010). She must depart the armchair to acquire enough experience as a participant-observer to gain competence in the practice. Thus, in writing about the nature of art, Timothy Binkley urged that 'what counts as a work of art must be discovered by examining the practice of art' (1977, 271) and Noël Carroll agreed that, 'a comprehensive theory of art must accommodate the facts as [the theorist] finds them revealed in our practices' (1999, 182). For these writers, working on theories of art particularly requires a familiarity with the dynamics of certain avant-garde art movements (see also Carroll 1993).

The idea is not that the probative judgements are judgements explicitly about the nature of the phenomenon. Rather, the method may be represented as a pattern of inferences. One begins with judgements that are characteristic of the relevant practice—for example, that such-and-such kind of work has a certain value. Added to this is a methodological principle, that the nature of a phenomenon is to be conceived so as to fit the judgements in question. From this, a conclusion is drawn about the nature of the phenomenon as it is embedded in the relevant practice.

An advantage of this variant over the armchair alternative is that it explains why it is reasonable to credit philosophers' judgements about cases. Judgements made by a philosopher who is knowledgeable about a practice conform to and hence indicate the norms that make the practice what it is and that determine the nature of the phenomenon under study. Notice that this is a special validation of the method, as it only applies to phenomena, including aesthetic and artistic phenomena, that are constituted (at least in part) through social practices.

The literature on the ontology of musical works represents the most sophisticated deployment of this method—perhaps because it has it that matters of some metaphysical weight turn on social facts. That is, a theory that states what it is for an item to be a musical

work takes it that a musical work is something that is implicated in musical practices, so that what is constitutive of the work is implicit in those practices. Jerrold Levinson's classic paper on 'What a Musical Work Is' (1980) claims that works of Western classical music are not eternally existing sonic patterns. The argument for this claim begins by considering judgements about the value of these works that appear to be sensitive to facts about the works' history of composition and prescribed performance means. Since timeless sonic patterns do not have histories or prescribed means of performance, these judgements speak against a Platonist ontology. Musical Platonists have replied by reinterpreting the same judgements as leaving open works' performance means and as consistent with the hypothesis that musical works are discovered and not created (e.g. Kivy 1983, 1987). Meanwhile, all parties agree that the debate only concerns Western classical music and that judgements of works in other musical traditions might implicate different conceptions of musical works (e.g. Davies 2001).

When practices are not as they appear to their participants, philosophers may seek to uncover the norms that make them up. A striking example is Frank Sibley's (2001) argument for the claim that visual art works are abstracta. This claim flies in the face of well-known arguments that works of visual art are, or are constituted by, physical objects because our practices generate judgements that seem to suggest that these works are forged by copying and destroyed by physical change (Goodman 1976, Wollheim 1980). Nevertheless, Sibley takes it for granted that 'it is to our practice that we must appeal. It would be improper to assert that a work of visual art is abstract if in our practice and dealings with it we clearly treat it as physical, and vice versa' (2001, 266). He then proceeds to cite evidence that we really do treat visual art works in ways that make sense only if they are abstract entities.

Practices need not be fixed, and we are sometimes convinced to try to change a practice because it is mistaken or in some other way open to criticism. In particular, the judgements that are characteristic of a practice might be inconsistent with each other and with some norms making up the practice. David Davies (2004, esp. 18–23) likens what happens in such a scenario to the process of reflective equilibrium, as understood by Goodman (1955) and Rawls (1971). This occurs gradually; it does not consist in the wholesale junking of the judgements generated in an artistic practice and the comprehensive rewriting of the norms making up that practice. Rather, judgements that are incompatible with existing norms are taken to provide sufficient reason to amend only those norms. The revision begins from within the practice itself, broadly conceived.

In Davies's hands, the process of reflective equilibrium is a variant on the method of conceptual analysis. The philosopher proposes theories that fit her judgements as a participant in a practice, but only in so far as her judgements cohere with norms of the practice that would survive rational reflection. In proposing a theory that fits some judgements, she may discount other judgements and advocate for certain norms as ones that should govern judgements in the practice. By the same token, her proposal may be rejected if it entails revisions to the practice that would not survive rational reflection. As Davies remarks, 'a theoretical account of our commerce with artworks stands in an essentially normative, and not merely descriptive, relation to the norms that operate in actual critical practice and the judgments in accordance with those norms that we actually make' (2004, 20).

Wholesale revisionism also has its fans, of course. Kendall Walton grants that analyses of concepts internal to an artistic or aesthetic practice may be good starting points. They are hypotheses, or 'candidates for acceptance by the philosopher but subject to rejection or modification, when the philosopher is interested in explaining the same body of data that the folk theory aims to explain' (Walton 2007, 155). Walton's groundbreaking theory of fiction illustrates how one might proceed when modification will not do and rejection is warranted. *Mimesis as Make-believe* opens with an assessment of philosophical work on fiction as 'so obviously in need of a fresh start that there can be no objection to my giving it one' (1990, 3). Walton subsequently introduces a completely technical concept of fiction that does not attempt to model folk concepts of fiction but that promises to explain a range of facts about appreciative engagement in a very wide range of contexts, some artistic, some not (see also Friend 2012).

Walton's theory of fictions as props in games of make-believe is revisionist in so far as it is a potential tool for thinking about fiction from an internal perspective, within our practices. His book is full of descriptions of our engagement with fictions that exploit his theoretical vocabulary so vividly and so insightfully that readers may begin to feel that vocabulary become their own. Thus Walton (2007) is careful to distinguish his approach from Goodman's, which adopts a forthrightly external perspective. Goodman warned the reader of *Languages of Art* that he 'must be prepared to find his convictions and his common sense—that repository of ancient error—often outraged by what he finds here. I have repeatedly had to assail authoritative current doctrine and fond prevailing faith.' (1976, xii).

The methods surveyed in this section have proven successful in attacking problems in aesthetics and the philosophy of art. Sometimes they have led to consensus. Sometimes they have framed persistent disagreements that have nevertheless yielded fruitful insights. Yet the track record is not perfect and there remain disagreements that have proven barren. Some of these disagreements are impasses.

An impasse is a disagreement that is barren because it is inextricably linked to methodological differences. Methodology is the science that concerns the appropriateness of methods for any given context of inquiry, and the choice of method is ideally common ground among inquirers. However, it may happen that parties to a debate advocate contrary theories precisely because they are impressed by different judgements about cases, where those judgements also determine the methods they use. The debate about theories of art is arguably at such an impasse (Lopes 2014a, ch. 3). Advocates of non-aesthetic theories of art appeal to judgements within our current practices that grant art status without regard to aesthetic considerations, and they take it to be a virtue of their theories that they model those judgements as is (e.g. Dickie 1984). Meanwhile, some of their opponents take it to be a virtue of aesthetic theories of art that they require some revision of our practices, possibly through a method of reflective equilibrium (e.g. Beardsley 1983). This is an impasse if the choice of judgements about cases determines the choice of method, which favours one type of theory or the other. When a disagreement concerns data, method, and results, there is no common ground.

The trouble is that one's choice of method may not be independent of one's substantive views. The questions one asks and the methods one selects to answer those questions may represent prior theoretical commitments. Such an impasse can only be cleared by appealing to considerations that are independent of the parties' philosophical stances.

5. Second-Order Methods

Critical demonstration and the variants on conceptual analysis described above are first-order methods. Philosophers employing these methods either engage in art criticism pregnant with philosophical implications or they make and appeal to judgements qua competent and reflective participants in an artistic or aesthetic practice. Philosophers using second-order methods do neither. Empirical art scholars in the humanities and social and behavioural sciences have developed powerful tools for gathering data, and philosophers may build theories given that data or the empirical hypotheses that explain that data.

The question of who is and is not a philosopher need not detain us. It is true, on one hand, that many art scholars in the humanities would be quick to point out that part of what they do is 'theory'. Aesthetics and the philosophy of art is a branch of philosophy, but it is not, as Walton writes, 'the private preserve of professional philosophers. Art historians, music theorists, and literary scholars frequently engage in philosophy, as do psychologists, cognitive scientists, and linguists' (2007, 147). It is also true, on the other hand, that so-called experimental philosophers borrow the methods of the empirical sciences, and some professional philosophers have ventured into experimental aesthetics (e.g. Meskin et al. 2013). To accommodate these complexities, we need only conceive empirical methods to be not distinctively philosophical. The data and empirical hypotheses that are products of these methods then supply material for second-order theorizing. Experimental philosophers and theoretically minded non-philosophers use both types of method in concert.

There are reasons to prefer second-order to first-order methods in aesthetics and philosophy of art. Critical demonstration puts philosophers in competition with humanistic art scholars, but few philosophers happen to have the training or temperament for writing successful art criticism. With a few exceptions, they are outclassed by scholars in departments of literature, music, fine arts, film, and theatre. A natural division of labour assigns philosophers the task of extracting from the best art scholarship such conclusions as may have some philosophical import. More importantly, there is reason to doubt that philosophers' judgements about cases provide good materials for conceptual analysis.

In the first place, empirical art scholarship sometimes provides debunking explanations of aesthetic and artistic phenomena. That is, it turns out that the factors driving our responses are not ones that we recognize from an internal perspective. On the contrary, they are sometimes ones that we will resist from an internal perspective. The methods of natural science are designed to protect against reinforcing our false beliefs, especially those held most dearly, and the same goes for the behavioural sciences, especially social psychology and sociology, as well as many branches of the contemporary humanities. The work of Pierre Bourdieu (1984) is a famous example of debunking research in sociology. Profuse examples of debunking hypotheses in humanistic art studies are provided by scholars working within Marxist, psychoanalytic, and feminist paradigms (e.g. Eagleton 1976, Mulvey 1989). Such hypotheses typically explain aesthetic or artistic response as functioning to sustain social structures in a way that clashes with how things appear from the internal perspective. Students of these approaches may find themselves unable to appreciate art works because they come to see them as subliminal instruments of social manipulation.

This kind of research outside philosophy poses challenges to philosophy done with first-order methods, but these challenges are not fundamental. As we have seen, debunking explanations from an external perspective do not demand revisionism from an internal perspective. Consider, for example, James Cutting's (2003) finding that mere exposure plays a surprisingly large role in determining aesthetic response. Mere exposure to a stimulus, which is perceived but not yet recognized, generally enhances the subject's attitude towards the stimulus (Zajonc 1968). Assuming that the more frequently images of paintings appear, the more likely it is that people will see them, Cutting measured preferences for Impressionist paintings and was able to draw several conclusions about the factors that determine variances in these preferences. Preference for a work is not explained by its canonicity, its prototypicality as an Impressionist work, or by subjects' expertise: it is largely a function of simple frequency of appearance. This is surprising, of course. However, as Cutting acknowledges, it does not mean that we should renounce our critical practices. The alternative is simply to admit that our preferences are often mistaken, by the standards of those practices. As Anthony Appiah puts it, 'where many people regularly make the wrong choice, a psychological explanation of why they do so supports the philosopher's claim that they are mistaken, because it relieves us of the worry that they are being guided by some rational principle we have failed to discover' (2008, 88). The same line may be taken in reply to debunking hypotheses from the humanities.

Whereas the challenge posed by Cutting's results is substantive and so may be met on substantive grounds, another body of empirical results puts in question the validity of the variants of conceptual analysis used in aesthetics and philosophy of art (Lopes 2014b). Richard Nisbett and Timothy Wilson's metastudy, 'Telling More Than We Can Know' (1977), is well known in philosophy for having planted doubts about the reliability of introspective access to mental states, and some of the studies surveyed by Nisbett and Wilson suggest that we have poor first-person access to the reasons for our aesthetic responses in particular. A more recent suite of studies by Petter Johansson and Lars Hall (esp. Johansson et al., 2005 and 2006) employ a powerful new protocol to investigate what they call 'choice blindness' (by analogy with change blindness). The protocol is to ask subjects to make an aesthetic choice in a forced-choice condition before asking them to state their reasons for their choice. The trick is that, through sleight of hand, half of the subjects are presented with the stimulus that they did not choose when providing their reasons. Few subjects notice the switch and it turns out that it makes no difference to the reasons they give whether they are presented with the stimulus that they chose or the one that they did not choose. Johansson and Hall conclude that reason-giving is post hoc confabulation. Another set of studies from Wilson's group indicate that subjects who are asked to reason about their aesthetic options make less good choices than do subjects who are asked simply to focus on their aesthetic options (e.g. Wilson and Schooler 1991). It appears that reasoning interferes with aesthetic choice.

These results compromise a methodological assumption needed to validate the use of the several variants of conceptual analysis described in the section 3.1. As we saw, philosophers' judgements as participant-observers in an aesthetic or artistic practice are taken as data for theory construction. The assumption is that critical judgements made by competent members of the practice carry information about the nature of the phenomena that are constitutive of the practice. The problem is that there is reason to believe that

some of these practices are not, contrary to appearances, practices of judgement. When critical judgements are made, they are made up, so that their content may tell us a great deal about our conception of a practice while telling us little about the practice itself.

As she resists the armchair variant of conceptual analysis in aesthetics and philosophy of art, Amie Thomasson writes that a theory is not established 'by little pictures or phrases explicitly entertained in the heads of competent speakers, but rather by the practices of those who use the terms and deal with the objects' (2010, 120–1). Although one might expect facts about any practice to be drawn from social science research, philosophers often rely on their own critical judgements as competent participants in the practice. There is now reason to believe that the methodological assumption permitting this short cut requires empirical validation for any given aesthetic or artistic subdomain. Note that this is not an instance of general scepticism about the validity of conceptual analysis in philosophy: it is a worry generated by the facts about the subject matter specific to one branch of philosophy.

These reflections use second-order method to bootstrap second-order methods. They do not compel us to scrap the traditional methods of the field: they advise only that those methods be supplemented by second-order methods. Moreover, it is left open what these methods may be and where they may be employed most effectively. The use of second-order methods is not yet mature enough for anyone to have anything more than hunches about these matters. Nevertheless, a start can be made by noting that second-order methods sort into two broad categories.

In the first, empirical data about aesthetic and artistic practices either supplant or accompany philosophers' own judgements when philosophical theories are formulated or verified. This approach is especially well represented in recent research on art emotion that leverages the results of emotion science (e.g. Robinson 2005, Bergeron and Lopes 2009, Coplan 2011). There is ample room for expansion. For example, most philosophers have been convinced by a few thought experiments that aesthetic judgements properly take account of contextual factors. A richer source of data may come from empirical research on the aesthetic dimension of the cognitive styles of humans in different cultural settings (e.g Masuda et al. 2008). Evolutionary psychology is another rich source of data that is only beginning to have an impact in the field (see Dutton 2008, Matthen 2013).

In the second category, philosophical theories are built to analyse the conceptual resources at work in the hypotheses proposed by empirical researchers to explain their own data. For example, an alternative to theories of art that negotiate philosophers' judgements about art is to fashion a theory of art that represents the role played by the concept of art in the explanatory hypotheses of historians, anthropologists, and other empirical art scholars (e.g. Moravcsik 1993, Lopes 2014a). On this model, aesthetics and the philosophy of art stands to the sciences of art as much philosophy of science stands to the sciences.

Aesthetics and the philosophy of art is a relatively young branch of analytic philosophy. Its problems were clearly identified only in the 1960s, principally by Beardsley, Goodman, Sibley, and Wollheim, and the methods used to address these problems continue to evolve. It is no accident that they are evolving as the arts evolve and as empirical studies of aesthetic and artistic phenomena become well established in the social and behavioural sciences as well as the humanities. As Wollheim advised, tact promotes neighbourly collaboration better than observing territorial boundaries.

References

Appiah, Anthony. 2008. *Experiments in Ethics*. Cambridge: Harvard University Press.

Batteux, Charles. 1746. *Les Beaux-arts Reduits à un Même Principe*. Paris: Durand.

Beardsley, Monroe. 1981. 'Aesthetic Experience', in *The Aesthetic Point of View*, ed. Michael Wreen and Donald Callen. Ithaca: Cornell University Press, pp. 285–97.

Beardsley, Monroe. 1983. 'An Aesthetic Definition of Art', in *What Is Art?* ed. Hugh Curtler. New Haven: Yale University Press, pp. 15–29.

Bergeron, Vincent and Dominic McIver Lopes. 2009. 'Hearing and Seeing Musical Expression', *Philosophy and Phenomenological Research* 78: 1–16.

Bergeron, Vincent and Dominic McIver Lopes. 2011. 'Aesthetic Theory and Aesthetic Science: Prospects for Integration', in *Aesthetic Science: Connecting Minds, Brains, and Experience*, ed. Steven Palmer and Arthur Shimamura. Oxford: Oxford University Press, pp. 63–79.

Binkley, Timothy. 1977. 'Piece: Contra Aesthetics', *Journal of Aesthetics and Art Criticism* 35: 265–77.

Bourdieu, Pierre. 1984. *Distinction: A Social Critique of the Judgment of Taste*, trans. Richard Nice. Cambridge: Harvard University Press.

Brady, Emily. 2003. *Aesthetics of the Natural Environment*. Edinburgh: Edinburgh University Press.

Brand, Peg Zeglin, ed. 2000. *Beauty Matters*. Bloomington: Indiana University Press.

Carlson, Allen. 2000. *Aesthetics and the Environment: The Appreciation of Nature, Art, and Architecture*. London: Routledge.

Carroll, Noël.1993. 'Historical Narratives and the Philosophy of Art', *Journal of Aesthetics and Art Criticism* 51: 313–26.

Carroll, Noël. 1999. *Philosophy of Art*. London: Routledge.

Coplan, Amy. 2011. 'Understanding Empathy: Its Features and Effects', in *Empathy: Philosophical and Psychological Perspectives*, ed. Amy Coplan and Peter Goldie. Oxford: Oxford University Press, pp. 3–18.

Currie, Gregory, Matthew Kieran, and Aaron Meskin, eds. 2014. *Aesthetics and the Sciences of Mind*. Oxford: Oxford University Press.

Cutting, James E. 2003. 'Gustave Caillebotte, French Impressionism, and Mere Exposure', *Psychonomic Bulletin and Review* 10: 319–43.

Danto, Arthur. 1986. *The Philosophical Disenfranchisement of Art*. New York: Columbia University Press.

Danto, Arthur. 1997. *After the End of Art: Contemporary Art and the Pale of History*. Princeton: Princeton University Press.

Davies, David. 2004. *Art as Performance*. Oxford: Blackwell.

Davies, Stephen. 2001. *Musical Works and Performances: A Philosophical Exploration*. Oxford: Oxford University Press.

Devereaux, Mary. 1995. 'Oppressive Texts, Resisting Readers, and the Gendered Spectator', in *Feminism and Tradition in Aesthetics*, ed. Peg Zeglin Brand and Carolyn Korsmeyer. University Park: Pennsylvania State University Press, pp. 121–41.

Dickie, George. 1962. 'Is Psychology Relevant to Aesthetics?' *Philosophical Review* 71: 285–302.

Dickie, George. 1984. *The Art Circle*. New York: Haven.

Dutton, Denis. 2008. *The Art Instinct: Pleasure, Beauty, and Human Evolution*. New York: Bloomsbury Press.

Eagleton, Terry. 1976. *Marxism and Literary Criticism*. Berkeley and Los Angeles: University of California Press.

Eaton, Anne. 2013. 'What's Wrong with the Female Nude?' in *Art and Pornography: Philosophical Essays*, ed. Jerrold Levinson and Hans Maes. Oxford: Oxford University Press, pp. 277–308.

Feagin, Susan. 1995. *Reading with Feeling: The Aesthetics of Appreciation*. Ithaca: Cornell University Press.

Friend, Stacie. 2012. 'Fiction as a Genre', *Proceedings of the Aristotelian Society* 112: 179–209.

Goodman, Nelson. 1955. *Fact, Fiction, and Forecast*. Cambridge: Harvard University Press.

Goodman, Nelson. 1976. *Languages of Art*, 2nd ed. Indianapolis: Hackett.

Isenberg, Arnold. 1949. 'Critical Communication', *Philosophical Review* 58: 330–44.

Johansson, Petter, Lars Hall, Sverker Sikström, Andreas Olsson. 2005. 'Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task', *Science* 310: 116–19 and supporting online material.

Johansson, Petter, Lars Hall, Sverker Sikström, Andreas Olsson. 2006. 'How Something Can Be Said about Telling More Than We Can Know: On Choice Blindness and Introspection', *Consciousness and Cognition* 15: 673–92.

Kivy, Peter. 1983. 'Platonism in Music: A Kind of Defense', *Grazer Philosophische Studien* 19: 109–29.

Kivy, Peter. 1987. 'Platonism in Music: Another Kind of Defense', *American Philosophical Quarterly* 24: 245–52.

Korsmeyer, Carolyn. 2002. *Making Sense of Taste: Food and Philosophy*. Ithaca: Cornell University Press.

Kristeller, P. O. 1951–2. 'The Modern System of the Arts', *Journal of the History of Ideas* 12: 496–527 and 13: 17–46.

Lamarque, Peter. 2005. 'Palaeolithic Cave Painting: A Test Case for Transcultural Aesthetics', in Thomas Heyd and John Clegg, eds, *Aesthetics and Rock Art*. Aldershot: Ashgate, pp. 21–36.

Levinson, Jerrold. 1980. 'What a Musical Work Is', *Journal of Philosophy* 77: 5–28.

Livingstone, Margaret. 2002. *Vision and Art: The Biology of Seeing*. New York: Abrams.

Lopes, Dominic McIver. 2014a. *Beyond Art*. Oxford: Oxford University Press.

Lopes, Dominic McIver. 2014b. 'Feckless Reason', in *Aesthetics and the Sciences of Mind*, ed. Gregory Currie, Matthew Kieran, and Aaron Meskin. Oxford: Oxford University Press, pp. 21–36.

Masuda, Takahiko, Richard Gonzalez, Letty Kwan, and Richard E. Nisbett. 2008. 'Culture and Aesthetic Preference: Comparing the Attention to Context of East Asians and European Americans', *Personality and Social Psychology Bulletin* 34: 1260–75.

Matthen, Mohan. 2013. 'Art and Evolution', in *Routledge Companion to Aesthetics*, 3rd edn., ed. Berys Gaut and Dominic McIver Lopes. London: Routledge, pp. 278–88.

Meskin, Aaron, Mark Phelan, Margaret Moore, and Matthew Kieran. 2013. 'Mere Exposure to Bad Art', *British Journal of Aesthetics* 53: 139–64.

Moravcsik, Julius. 1993. 'Why Philosophy of Art in Cross-Cultural Perspective?' *Journal of Aesthetics and Art Criticism* 51: 425–35.

Mulvey, Laura. 1989. *Visual and Other Pleasures*. London: Macmillan.

Nehamas, Alexander. 2007. *Only a Promise of Happiness: The Place of Beauty in a World of Art*. Princeton: Princeton University Press.

Nisbett, Richard and Timothy Wilson. 1977. 'Telling More Than We Can Know: Verbal Reports on Mental Processes', *Psychological Review* 84: 231–59.

Nussbaum, Martha. 1992. *Love's Knowledge: Essays on Philosophy and Literature*. Oxford: Oxford University Press.

Palmer, Steven and Arthur Shimamura. 2011. *Aesthetic Science: Connecting Minds, Brains, and Experience*. Oxford: Oxford University Press.

Rawls, John. 1971. *A Theory of Justice*. Cambridge: Harvard University Press.

Robinson, Jenefer. 2005. *Deeper Than Reason: Emotion and Its Role in Literature, Music, and Art*. Oxford: Oxford University Press.

Saito, Yuriko. 2008. *Everyday Aesthetics*. Oxford: Oxford University Press.

Schaeffer, Jean-Marie. 2000. *Art of the Modern Age: Philosophy of Art from Kant to Heidegger*, trans. Steven Rendall. Princeton: Princeton University Press.

Schellekens, Elisabeth. 2011. 'Experiencing the Aesthetic: Kantian Autonomy or Evolutionary Biology?' in *The Aesthetic Mind: Philosophy and Psychology*, ed. Elisabeth Schellekens and Peter Goldie. Oxford: Oxford University Press, pp. 223–38.

Shiner, Larry. 2001. *The Invention of Art: A Cultural History*. Chicago: University of Chicago Press.

Sibley, Frank. 2001. 'Why the *Mona Lisa* May Not Be a Painting', in *Approach to Aesthetics: Collected Papers on Philosophical Aesthetics*, ed. John Benson, Betty Redfern, and Jeremy Roxbee Cox. Oxford: Oxford University Press, pp. 256–71.

Thomasson, Amie. 2005. 'The Ontology of Art and Knowledge in Aesthetics', *Journal of Aesthetics and Art Criticism* 63: 221–9.

Thomasson, Amie. 2010. 'Ontological Innovation in Art', *Journal of Aesthetics and Art Criticism* 68: 119–30.

Walton, Kendall. 1990. *Mimesis as Make-Believe: On the Foundations of the Representational Arts*. Cambridge: Harvard University Press.

Walton, Kendall. 2007. 'Aesthetics—What? Why? and Wherefore?' *Journal of Aesthetics and Art Criticism* 65: 147–62.

Wilson, Timothy D. and Jonathan W. Schooler. 1991. 'Thinking Too Much: Introspection Can Reduce the Quality of Preferences and Decisions', *Journal of Personality and Social Psychology* 60: 181–92.

Wollheim, Richard. 1980. *Art and Its Objects*, 2nd ed. Cambridge: Cambridge University Press.

Wollheim, Richard. 1999. *On the Emotions*. New Haven: Yale University Press.

Zajonc, Robert B. 1968. 'Attitudinal Effects of Mere Exposure', *Journal of Personality and Social Psychology Monographs* 9: 1–27.

Zangwill, Nick. 1995. 'Groundrules in the Philosophy of Art', *Philosophy* 70: 533–44.

Zeki, Semir. 2000. *Inner Vision: An Exploration of Art and the Brain*. Oxford: Oxford University Press.

CHAPTER 34

THE METHODOLOGY OF LEGAL PHILOSOPHY

ALEX LANGLINAIS AND BRIAN LEITER

1. Introduction

Legal philosophy,[1] certainly in the anglophone world and increasingly outside it, has been dominated for more than a half-century by H. L. A. Hart's 1961 book *The Concept of Law* (Hart 1994). Unsurprisingly, then, methodological debates in legal philosophy typically scrutinize either one of two (related) methodological claims in Hart's classic work. The first is that his theory is both general and descriptive (Hart 1994: 239). The second is that his theory is an exercise in both linguistic analysis and descriptive sociology (Hart 1994: vi). What do these two claims reveal about Hart's theoretical ambitions and methodological commitments, and what light do they shed on the debates in legal philosophy since then?

2. Hart's Methodology

In the Postscript to *The Concept of Law*, Hart says that his theory is *general* in that its aim is to give an account of the general structure and features exhibited by all modern legal systems, rather than the idiosyncratic features of any particular legal system. (In this respect, it is different than Ronald Dworkin's theory, which on Hart's [plausible] view is a case of "particular" jurisprudence, that is, an account of the law of particular jurisdictions, such as the Anglo-American ones.) He writes:

> It is *general* in that it is not tied to any particular legal system or legal culture, but seeks to give an explanatory and clarifying account of law as a complex social and political

[1] Alex Langlinais is a candidate for the J.D. at Yale University; Brian Leiter is Karl N. Llewellyn Professor of Jurisprudence and Director of the Center for Law, Philosophy and Human Values at the University of Chicago.

> institution with a rule governed (and in that sense 'normative') aspect. This institution, in spite of many variations in different cultures and in different times, has taken the same general form and structure, though many misunderstandings and obscuring myths, calling for clarification, have clustered round it.
>
> (Hart 1994: 239–40)

Hart's theory is *descriptive* in that his attempt to explain and clarify the general features of modern legal systems includes no assessment of the (moral) value or reason-giving force of legal rules or legal systems in general or in any particular case. "My account is *descriptive*," Hart says, "in that it is morally neutral and has no justificatory aims: it does not seek to justify or commend on moral or other grounds the forms and structures which appear in my general account of law" (Hart 1994: 240). On Hart's view, we can give an adequate constitutive explanation of the general features of legal norms and systems without considering the moral value or hazards of having law, whether and when one is obligated to obey the law, and so on. Whether a theory of law can be both general and descriptive in this sense is one point of contention in legal philosophy to which we will return, but for now we can just focus on what this claim says about Hart's theoretical ambitions.

The object of Hart's theory is the complex social institution we call a legal system, one that he thinks can be illuminated by examining our ordinary ways of talking about laws and legal systems (about which more later in this section). The most important elements of Hart's theory are well known to legal theorists. First, law is a union of primary and secondary rules (a combination of rules that tell ordinary citizens what they can and cannot do, and rules *about* all the other rules, e.g. rules of adjudication and legislation) (Hart 1994: 80–1). Second, in every legal system, there is a secondary rule that Hart calls the "rule of recognition" which specifies the criteria any norm must satisfy to count as a valid legal norm of that system (Hart 1994: 94–5, 103).[2] Third, the rule of recognition in any legal system is, by contrast to all the other norms, simply a social rule constituted by the convergent practice of its legal officials in applying certain criteria of legal validity which they also view themselves as having an obligation to apply (Hart 1994: 101–2, 110); all other norms are legally valid in virtue of satisfying the criteria in the rule of recognition. Finally, on Hart's account, it is not necessary for a norm to be legally valid that it exhibit any moral feature or set of features (Hart 1994: 185–6). These claims are meant to apply to modern legal systems and to give a unifying explanation of what the ordinary person knows about them. Do they, then, define "the essence" of a legal system, or does Hart implicitly reject an essentialist picture?

Leslie Green (1996) argues that Hart's recognition of borderline cases of law implies that his theory is an anti-essentialist one. Given that our concept of law or any revisionary concept we might adopt will inevitably confront borderline cases, it follows, he says, that "[f]or Hart, there is no essence to the phenomena we call 'law'" (Green 1996: 1692). There may of course be properties *essential* to laws or legal systems even if there are no necessary and sufficient conditions. Hart claims, for example, that any society with a legal system must have a rule of recognition. Although he does not think the

[2] Hart and others think there can be more than one rule of recognition in a legal system, but for sake of elegance, we will speak of "rule of recognition" in the singular in what follows.

existence of such a rule is sufficient for a legal system: it is essential. Admittedly, Hart was deeply influenced by a tradition in post-World War II anglophone philosophy associated with J. L. Austin, Gilbert Ryle, and the later Wittgenstein that was uneasy with the whole idea of essential properties and suspicious of metaphysics in general. An anti-essentialist approach to a historically contingent social institution like law certainly would have been in keeping with Hart's time and philosophical background. Moreover, *The Concept of Law* often displays a rhetorical restraint that suggests this was just the approach Hart intended.

Still, the balance of evidence supports a reading of Hart as a modest (if reluctant) essentialist. Otherwise it would be hard to explain what Hart says, for example, about the existence conditions of legal systems:

> There are therefore two minimum conditions necessary and sufficient for the existence of a legal system. On the one hand, those rules of behavior which are valid according to the system's ultimate criteria of validity must be generally obeyed, and, on the other hand, its rule of recognition specifying the criteria of legal validity . . . must be effectively accepted as common public standards of official behavior by its officials.
>
> (Hart 1994: 116)

The passage admits, it seems, of only an essentialist reading.[3] The existence of borderline cases, even in light of passages like the preceding, might just signal our inability to settle certain evidentiary questions: that we do not *know* where to locate the borderline cases does not mean there is no fact of the matter about where the border lies. We will proceed, then, on the assumption that Hart's theoretical ambitions are to explain and clarify the general features of modern legal systems, at least some of which rise to the level of essential properties. And this is certainly how he has been understood by his followers. Julie Dickson, for example, following Hart's student Joseph Raz, says that,

> A successful theory of law . . . is a theory which consists of propositions about the law which (1) are necessarily true, and (2) adequately explain the nature of law. . . . I am using "the nature of law" to refer to those essential properties which a given set of phenomena must exhibit in order to be law.
>
> (Dickson 2001: 17)

She is echoed more recently by Scott Shapiro, who says that in inquiring into "the fundamental nature of law" we want to "supply the set of properties that make (possible or actual) instances of [law] the things that they are"(2011: 8–9) and offers the example of water being H_2O: "Being H_2O is what makes water *water*. With respect to law, accordingly, to answer the question 'What is law?' on this interpretation is to discover what makes all and only instances of law instances of *law* and not something else" (2011: 8–9).[4] In addition, says Shapiro (here again echoing Dickson who is following Raz), "to discover the law's nature" is also "to discover its necessary properties, i.e., those properties that law could not fail to have" (2011: 9).

[3] This would be compatible with Hart thinking that the concept of "law" has no essence, even if the concept of a "legal system" does.

[4] This is a surprisingly strong claim for many reasons, not least of which is that water is a natural kind, while law is an artifact.

How does Hart propose to fulfill the ambitions of producing a theory that is both general and descriptive? This brings us to his claim to be engaged in both linguistic analysis and descriptive sociology. *The Concept of Law* is not a piece of empirical social science, as any quick skim through its pages would confirm. All Hart means when he says his theory is a piece of descriptive sociology is that his goal is to give an adequate account of a particular kind of social institution, namely law. But how does linguistic analysis fit into the picture, and how does it relate to the descriptive sociological project? The type of linguistic analysis Hart has in mind is roughly the type associated with the ordinary language philosophers of post-war Oxford University, though Hart's claim to be engaged in this kind of project can be somewhat misleading. Unlike the ordinary language philosophers, Hart rarely argues explicitly for a position on the basis of some observations about the appropriateness of various expressions for a given context.[5]

However, his method does reflect the overall *spirit* of ordinary language philosophy: that we can gain philosophical insight about some phenomenon by attending to the conceptual distinctions we use to talk and think about it. In the preface to the *The Concept of Law*, Hart characterizes the relationship between this type of linguistic analysis and descriptive sociology as follows:

> Notwithstanding its concern with analysis the book may also be regarded as an essay in descriptive sociology; for the suggestion that inquiries into the meanings of words merely throw light on words is false. Many important distinctions, which are not immediately obvious, between types of social situation or relationships may best be brought to light by an examination of the standard uses of the relevant expressions and of the way in which these depend on a social context, itself often left unstated. *In this field of study it is particularly true that we may use, as Professor J.L. Austin said, 'a sharpened awareness of words to sharpen our perception of the phenomena'.*
>
> (Hart 1994: vi, emphasis added)

The key to understanding Hart's method—and thus its influence on subsequent legal philosophy—is to understand the precise sense in which we can use a sharpened awareness of words to sharpen our awareness of legal phenomena. Hart was convinced that it is especially true of social phenomena like law that we can employ linguistic analysis to understand their nature. On its face, this seems puzzling: why should the actual nature of a social phenomenon be hostage to how ordinary people talk? A short detour into the metaphysics of social reality may provide a rationale, though not the one Hart had in mind.

For Hart and all legal positivists, legal systems are *socially constructed* (cf. Green 1996: 1687, 1691–2). To understand the philosophical import of that claim, it will prove helpful to introduce some ideas from John Searle's influential work on the "construction of social reality" as he calls it (Searle 1995, 2009). Searle's work obviously postdates Hart's, but it was deeply influenced by the same philosophical climate of ordinary language

[5] There are some exceptions: Hart's discussions of the difference between "being obliged" and "having an obligation", and between "nullity" and "sanction" are cases where the book's implicit method becomes explicit (see Hart 1994: 82–3, 33–5). These arguments are quite important, since they are supposed to establish one of Hart's major disagreements with earlier positivists like John Austin and Hans Kelsen: against them, Hart argues, through these ordinary language examples, that there is no necessary connection between law and coercion.

philosophy that influenced Hart. To say that law is socially constructed is to say that "legal facts"—for example, that Buick is liable to MacPherson for its negligent design of the car,[6] that the speed limit on Lakeshore Drive in Chicago is 45 mph, or that Antonin Scalia[7] is an associate justice of the Supreme Court of the United States—are certain kinds of *socially constructed facts*.[8] Unlike physical facts, these kinds of social constructs are realizable only in the context of human institutions (Searle 1995: 2). Take Searle's frequent example, money. It is not a fact about the natural world that a greenish piece of paper with a picture of George Washington has a value of one dollar throughout the United States. Rather, that these physical facts constitute the social construct *of being a dollar* depends on the fact that those physical facts are so recognized by members of American society.

It is, thus, a hallmark of social constructs that they exist if and only if a society—or some subset of that society—collectively recognizes them as existing (Searle 1995: 32–4, 52–3). That is to say, these kinds of social constructs are constituted and sustained by a kind of collective agreement—at least, in the case of law, an "agreement" among those Hart calls "officials" of the system. Scalia is an associate justice of the U.S. Supreme Court because *and only because* other officials of the American legal system recognize that he is one and treat him accordingly (we shall say more about this in this chapter). Although such facts are socially constructed, it is still possible for any individual to be mistaken about them. One can be mistaken, for example, about who is an Associate Justice of the U.S. Supreme Court, because whether or not any particular individual is on the Court is not a matter of anyone's individual opinion. So, too, any individual can mistake a one-dollar American bill for a ten-dollar bill. Social constructs are not individual constructs: it is not up to any individual *what an American one-dollar bill is*, and it is not up to any individual *who is an American Supreme Court justice*.

Social constructs generally involve the imposition of a status or function on a person, object, or event that it could not have solely in virtue of its physical features (Searle 1995: 39–42, 2009: 7). Scalia, for example, is not a Supreme Court justice just in virtue of the physical facts about his person. Rather, Scalia is a justice because we impute to him the status of a particular kind of judge, with all of the powers and duties that entails. More precisely, on Hart's view, Scalia is an associate justice of the U.S. Supreme Court because officials of the American legal system recognize that he satisfied the criteria for occupying such a role that are set out by the public power-conferring rules (partly specified in the U.S. Constitution) for becoming a justice of the Supreme Court. The imposition of these statuses or functions is accomplished through collective recognition of rules of the form "X counts as Y in context C," where the X term specifies some physical object or objects and the Y term specifies some status or function that things satisfying the X term cannot have simply in virtue of their satisfaction of the X term. These involve what Searle calls *constitutive rules* (Searle

[6] *MacPherson v. Buick Motor Co.,* 217 N.Y. 382, 111 N.E. 1050 (1916).

[7] Scalia, born in 1936, has been an influential conservative Supreme Court justice in the United States since 1986. Nothing turns on our choice of example, we have simply picked a jurist well known in our jurisdiction.

[8] We are using slightly different terminology than Searle. We use "social construct" to refer to aspects of reality that are constituted by human beliefs and attitudes; we use "social facts" as it is usually used in the jurisprudential literature to refer to those beliefs and actions of officials of a legal system that are constitutive of law.

1995: 43–9; 2009: 9–10, 96–8), and they play a particularly important role in Hart's theory. As we mentioned above, one of Hart's most important contributions to legal philosophy is the idea that a norm is legally valid in a given legal system if and only if it satisfies the constitutive criteria specified by the rule of recognition of that system. But a rule of recognition for Hart is also just constituted by the actual practices and attitudes of legal officials in deciding questions of legal validity. That is, on Hart's view, the rule of recognition has the content it does (e.g. "no norm is legally valid unless enacted by Parliament") *only* in virtue of the fact that officials, in fact, treat enactment by Parliament as a criterion of legal validity and the officials take themselves to have an obligation to do so. So a rule of recognition is itself a constitutive rule, and it is one that is constituted by the acts and attitudes of legal officials. A legal system exists, at bottom, because of the actions and beliefs of officials of that system: their practices and attitudes *constitute* the legal system.

Law, then, is an example of a social construct in something like Searle's sense. Yet as Searle also notes, "language is essentially constitutive of institutional reality [i.e. social constructs]" (Searle 1995: 59). If this is right, it follows that linguistic analysis is an important if not indispensable part of legal theory. We can and should use a sharpened awareness of words (construed broadly to include not just words but also whole expressions, speech acts,[9] and the mental states revealed therein) to sharpen our perception of legal phenomena because these phenomena are wholly constituted by the beliefs and attitudes these linguistic practices reveal. And so it appears that Hart's theory can be both linguistic analysis and descriptive sociology in one stroke. If law is socially constructed, and social constructs are constituted by language, then when we analyze our ordinary ways of speaking and thinking about legal systems, we are simultaneously analyzing the general features of a very important human institution.

This basic methodological position would explain many of Hart's most important arguments in *The Concept of Law*. Consider Hart's distinction between *duty-imposing* and *power-conferring* rules. Duty-imposing rules obligate others to do or refrain from doing something, while power-conferring rules specify procedures through which certain individuals can acquire legal powers and thus alter the duties of others and/or themselves. Hart's case against command theories of law (according to which laws are the commands of a sovereign backed by threats of sanction) relies heavily on the claim that we need the concept of a power-conferring rule to explain various legal phenomena (rights, wills, contracts, etc.) and that command theories cannot explain such rules without distortion (Hart 1994: 40–42, 48–49). The distinction is also called on to explain the nature of secondary rules of change and adjudication. Hart's argument for this distinction is simply this:

> Such power-conferring rules are thought of, spoken of, and used in social life [via speech acts] differently from rules which impose duties, and they are valued for different reasons. What other tests for difference in character could there be?
>
> (Hart 1994: 41)

In other words, all we need to know in order to distinguish these two types of rules is that they play different roles in social life, and the fact that they play such different roles

[9] The relevant "speech acts" will include those that involve "overruling," "deciding for the plaintiff," "awarding injunctive relief," and so on.

is exhausted by the way we conceptualize and talk about those roles. Our "talk" and the behavior that accompanies that talk *is the evidence* of the differing social roles and indeed explains the distinction itself.

Hart's method can now be summarized by the following claims: (1) a theory of law should be an account of the essential properties of modern legal systems in general; (2) a theoretical account of law's essential properties can proceed without consideration of its value or reason-giving force; (3) legal systems are socially constructed, and a theory of law is therefore a theory of the social facts about beliefs, attitudes, and actions that constitute a legal system; (4) this social construct is amenable to linguistic analysis, given that it is constituted by language; and (5) given that law is socially constructed, a general theory of law is just an attempt to elucidate the folk concept of law, that is, the concept manifest in the language we use to think and talk about it.

3. Objections to Hart's Methodology

We shall consider in this section several lines of objection to Hart's influential approach to the problems of legal philosophy. One line of objection questions the fruitfulness of the methodology of conceptual analysis, as described above. Another line of objection largely accepts the method of "linguistic" or "conceptual" analysis, but questions whether there is a concept of law there that can simply be "described." We take these up in turn.

3.1 The Naturalistic Challenge

In many areas of philosophy, doubts about the kind of conceptual and linguistic analysis Hart relies upon have become common (e.g. Weinberg et al. 2001), but not so in legal philosophy, where almost everyone, following Hart, employs the method of appealing to intuitions about possible cases to fix the referent of "law," "legal system," "authority," and the other concepts that typically interest legal philosophers. The Hartian approach to jurisprudence thus seems to be a case of what Frank Jackson (1998: 43) calls "immodest" conceptual analysis, in the sense that it aims to deliver knowledge about the actual nature of its subject-matter by consulting intuitions about how the concept extends to possible cases (the contrast is with conceptual analysis that aims only to determine *to what* the concept refers, while allowing that the sciences will answer the question about the actual character of the referent, if there is one). Hart identifies the touchstone of his theory as "the widespread common knowledge of the salient features of a modern municipal legal system which … I attribute to any educated man" (Hart 1994: 240). Any educated person in a modern society, Hart thinks, knows a fair bit about legal systems. Most importantly, Hart thinks any educated person can be expected to have the following skeletal understanding of these general features of law:

> They comprise (i) rules forbidding or enjoining certain types of behavior under penalty; (ii) rules requiring people to compensate those whom they injure in certain ways; (iii) rules specifying what must be done to make wills, contracts or other arrangements which confer

> rights and create obligations; (iv) courts to determine what the rules are and when they have been broken, and to fix the punishment or compensation to be paid; (v) a legislature to make new rules and abolish old ones.
>
> (Hart 1994: 3)

In other words, we can expect the (educated) folk concept of law to identify these as the real features of legal systems. Hart's method is, thus, clearly a form of immodest conceptual analysis in Jackson's sense, but (and this is key) it has to be since the concept, as manifest in our language, constitutes the social construct of law! Hart's theory may rely somewhat less obviously on intuitions about possible cases than many contemporary legal philosophers do (like Joseph Raz, Matthew Kramer, and Scott Shapiro), but it is still the case that on Hart's theory we can derive knowledge about the actual nature of law by analyzing our shared concept of it as manifest in ordinary language.

Of course, there is more to elucidating a folk concept than opinion polling or intuition pumping, and there has to be conceptual space for revision of the concept. Folk reports about some phenomenon are not infallible guides even to the folk concept of that phenomenon (and for obvious reasons: they can be unreflective, inconsistent, etc.). As Searle points out, a society can go about creating and sustaining institutions without being aware that this is what they are doing (Searle 1995: 47). In particular, the members of a society can fail to be conscious of the patterns of collective agreement that constitute their own institutions. This could be because social institutions tend to evolve over time, such that there is no discernible point in a society's history when its members engage in any overt act of social construction. A society might also fail to be conscious of these collective agreements because they do not even realize that some institution of theirs is in fact socially constructed (gender roles are a relevant example here). Nevertheless, these collective agreements are an important part of the folk concept, and it is the theorist's job to bring them to light. We find that this is the best way to understand Hart's introduction of the concept of a rule of recognition. The idea that what makes any norm legally valid in a given legal system is its satisfaction of the criteria specified by a particular social rule (namely, the rule of recognition) is not intuitively obvious to the ordinary person. Indeed, it is not even intuitively obvious to legal professionals or many legal philosophers! But the idea of a rule of recognition is implicit in our folk concept of law—in the fact that everyone distinguishes, for example, between legally binding norms and other norms, and that criteria of legal validity can and do differ between legal systems—and introducing it into one's theory is necessary to make sense of other familiar features captured by the folk concept according to Hart.

Conversely, Hart aims to rid the folk concept of certain inconsistencies; in this respect, his kind of conceptual analysis is explicitly revisionary. For example, it certainly seems to be part of the folk understanding of law that legal systems guide conduct primarily through coercion or the threat of coercion (indeed, Hans Kelsen, the other great theorist in the modern positivist tradition, thought coercion was essential to law!). Hart was surely aware of the prevalence of this belief, but he is explicit that it is mistaken:

> Plainly we shall conceal the characteristic way in which such rules function if we concentrate on, or make primary, the rules requiring courts to impose the sanctions in the event of disobedience; for these latter rules make provision for the breakdown or failure of the primary purpose of the legal system. They may indeed be indispensable but they are ancillary.
>
> (Hart 1994: 39)

In other words, while the use of coercion is an important feature of legal systems (bordering on the indispensable), we will fail to do justice to other important features of the folk concept—such as the idea that law can impose *obligations*—if we overstate its role. So, for Hart, the legal theorist's task is to elucidate the folk concept of law, but this involves throwing light on implicit features of our concept that may go unnoticed by the ordinary person.

Ordinarily, naturalists worry that "folk" intuitions about the extensions of concepts cannot be informative as to the actual nature of their referents since what the folk believe is hostage to ignorance and other epistemic infirmities. Naturalists typically defer to the more epistemically robust methods of the various sciences: if our best physics says that space can be non-Euclidean, then Kant's "a priori" intuition about the structure of space be damned! Why not think the same is true of law? Why not defer to our best social scientific theory of law and legal phenomena to demarcate what law really is? Hart obviously does not adopt that approach: he treats ordinary language and concepts as his starting point, and his revisions appeal only to inconsistencies and tensions that are manifest in how the "folk" think and talk about law.

One difficulty for the naturalist is that there is not, at present, an epistemically robust social science of law and legal phenomena (cf. Leiter 2007: 192). But Joseph Raz has given a different, and more ambitious, answer to the naturalist challenge:

> In large measure what we study when we study the nature of law is the nature of our own understanding. The identification of a certain social institution as law is not introduced by sociologists, political scientists, or some other academics as part of their study of society. It is part of the self-consciousness of our society to see certain institutions as legal. And that consciousness is part of what we study when we inquire about the nature of law.
>
> (Raz 2009a: 31)

This rejoinder to the naturalist, however, seems weaker than Hart's position actually allows. It is not just that law "is part of the self-consciousness of our society to see certain institutions as legal," it is that—per Hart's actual method and Searle's related account of socially constructed reality—that law *really is* what society, or some subset of society ("officials" of the system in Hart's terminology), "understands" it to be, perhaps not self-consciously of course, but in terms of their practices and attitudes, which are made manifest in their language.

Again, consider the case of money. In fact, money is a concept of interest to sociologists, political scientists, and other scholars who study social and economic orders. But the fact that some piece of metal or paper *is money* (such that it admits of empirical study) is conceptually prior to the claims of sociologists and political scientists, and is, per Searle, constituted by the attitudes of people in that society. If law is the same, then the stronger argument against the naturalist is that the metaphysics of a social construct like law *precludes* deference to the empirical sciences for purposes of general jurisprudence, since the way people use and understand the concept just constitutes the fact in question. Interestingly, this "metaphysical" defense of Hart's methodology—which is the methodology of almost all legal philosophy these days—is not one offered by Hart himself or any of his many followers. Yet it may be the best way to resist the familiar kind of naturalistic worries about "immodest" conceptual analysis (for such worries in the jurisprudential context, see Leiter 2007: 175–81, 183–99).

Yet this kind of argument does not wholly deflect the naturalist's challenge. For a naturalist can reasonably allow that a social phenomenon like law is, at least in the first instance, individuated by shared beliefs and attitudes, and still press the point that the "folk" concept of the phenomenon ought to be revised in light of whatever refined understanding of the phenomenon is explanatorily and predictively fruitful. After all, even the proponent of Hart's basic methodological posture allows that the understanding of the ordinary "folk" needs revision in various ways, as long as it does not forego some core of folk commitments. The Razian rejoinder, then, that we want to "understand our own self-understanding" is neither here nor there—the latter is a fine topic for social psychology and anthropology, but if it turns out that sacrificing parts of the folk concept yields a more powerful theoretical understanding of law as a phenomenon in human societies, why should we prefer Razian hermeneutics? If, in fact, the social-scientific accounts of law and legal phenomena rise to the level that warrants epistemic confidence, why not revise the folk concept so that it fits with what the sciences discover? A different kind of philosophical naturalist, the experimental philosopher, can also properly object at this point that if theorists *really* want to "understand our own self-understanding," they ought to adduce reliable evidence of what that understanding is rather than rely on the armchair method of appeal to intuitions about how the concept applies in possible cases.

3.2 Objections to "Descriptive" Jurisprudence

Even those comfortable with Hart's "armchair" methods have expressed doubts about Hart's evident assumption that there is in fact a robust and homogeneous folk concept of law for the theorist to describe: a cohesive set of beliefs, attitudes, and dispositions that can be found in every modern society that has law. Hart never offers much motivation for this assumption in his published work, and the second line of objection we will consider here suggests that his optimism is unwarranted. This objection, put forward first in 1980 by John Finnis (Finnis 2011) and then in a series of papers by Stephen Perry (e.g. Perry 2001), argues that every theory has to select the important features of its subject-matter that merit theoretical accounting but that there is no way to do that without first engaging in moral reflection on the value of law and legal systems. These critics thus reject Hart's idea that we can explain what law is without any consideration of its moral value or reason-giving potential.[10] Whereas the methodological debates in epistemology, philosophy of mind, and other

[10] It is, of course, not the only possible motivation for rejecting Hart's descriptive aspirations. One could accept that there is a shared folk concept of law *and* that the goal of legal theory is to describe that concept while also rejecting the idea that legal theory can be descriptive. For example, one could argue that legal systems have an intrinsic moral objective or function and that we cannot therefore understand the nature of law without reference to this objective or function. On this view, part of the legal theorist's task is to describe the conditions under which law succeeds or fails to achieve its objective or discharge its function, and this task unquestionably involves moral argument. All the same, one could think that this function or point is implicit in the relevant social facts and that it is the legal theorist's job to bring this to light. Examples of such approaches include Murphy (2006) and, arguably, Shapiro (2011). The viability of this view turns on many substantive questions of legal philosophy that we cannot address here, but it is worth noting this point in logical space.

areas have taken for granted that the aim is to analyze and describe the contours of the concepts at issue,[11] in legal philosophy the main debate, perhaps surprisingly, has been about whether a good theory of its subject matter can proceed without any substantive moral theorizing. And it would not be an overstatement to say that the literature on methodological issues in legal philosophy has been primarily concerned with this question.

More precisely, *methodological positivists* claim that a descriptive account of law's nature is explanatorily prior to any moral assessment or evaluation of law, including any moral assessment or evaluation of whether law provides us with genuine reasons for action. *Methodological anti-positivists* deny the preceding: a theory of law does require moral theorizing because we simply cannot understand what law is prior to some account of what law ought to be.

There is no master argument for methodological positivism. It is a hypothesis about how parsimonious a theory of law can be, and it is supported to the extent that an explanatorily adequate theory can be given consistent with that hypothesis. Methodological anti-positivism on the other hand must be motivated by some perceived shortcoming in positivist theories of law, and the present line of objection attempts to locate one.

Finnis's version of the objection begins with an innocuous observation about theory construction: a theory of some phenomenon must begin with some assumptions about which aspects of it are significant or important for the purposes of explanation. Without some provisional discrimination between the important and unimportant features of a phenomenon, what we get is not a theory at all but "a vast rubbish heap of miscellaneous facts" (Finnis 2011: 17). Finnis's question is this: how can we identify the important features of law in a principled way?

We have already surveyed Hart's answer to that question. The salient features of law are the ones picked out as salient by the folk concept. Since the folk concept of law constitutes the phenomenon to be explained, this is as principled as a judgment of salience could hope to be. The theorist does not have to accept all the folk explanations of these features, because these can fail to track the folk concept for the reasons we have mentioned. But descriptive adequacy to what the "ordinary man" knows about law is the starting point for the theory.

Finnis thinks this answer fails because there is no folk concept whose judgments can be authoritative with respect to all legal systems. He agrees with Hart that legal systems are constituted by various facts about the behavior and attitudes of officials and that the legal theorist's task is to explain these patterns. But to this Finnis adds an important proviso: we cannot fully understand these social facts without understanding "their objective, their value, their significance or importance, *as conceived by the people who performed them, engaged in them, etc.*" (Finnis 2011: 3, emphasis added; Finnis 2003: 118). In short, we cannot fully understand a society's legal practices without accounting for the participants' understanding of the point or purpose of those practices. (Hart denies this, but we put that to one side for now.) There is, however, no universal or even general agreement about the point or purpose of law. Societies can and do vary greatly in their conceptions of this point or purpose (Finnis 2011: 4), and this variation will manifest itself as significant variation in the social facts that constitute these societies' respective legal systems.

[11] Radical naturalists like Hilary Kornblith go further, denying that concepts are even what we are interested in, as opposed to the phenomena themselves.

If that is right, then there is no one folk concept of law shared by every society that has law. There are only particular, and perhaps incommensurable, conceptions of law. "*How then*" Finnis asks, "*is there to be a general, descriptive theory of these varying particulars?*" (Finnis 2011: 4). Analyzing our folk concept clearly will not do, because that will only get us an account of some particular conception of law or a loose assemblage of different conceptions. "And jurisprudence, like other social sciences, aspires to be more than a conjunction of lexicography with local history, or even the juxtaposition of all lexicographies conjoined with all local histories" (Finnis 2011: 4).

Finnis argues the only solution is to adopt a *central case methodology*, a method that takes its cue from Aristotle's notion of focal meaning or core-dependent homonymy (Finnis 2011: 9–11, 429–30). A central case analysis of some phenomenon identifies some subset of possible or actual instances of that phenomenon as explanatorily privileged. The members of this subset are the paradigm or *central cases* of the phenomenon, and they are privileged in two respects. First, the central cases are privileged insofar as a theory of the phenomenon is primarily concerned with explaining the important features of these cases. Second, the central cases are explanatorily prior to those instances of the phenomenon that are not members of the set of central cases. These are the *peripheral cases* of the phenomenon, and they are explanatorily posterior in that they can only be understood as defective, failed, or "watered-down" (Finnis 2011: 11) versions of the central cases.

How do we identify the central case of law? Finnis's answer is that the central case of a legal system is one whose valid norms are at least presumptive moral requirements, and that it is therefore impossible for the theorist to describe the central cases of law without engaging in some moral theorizing to determine what would have to be true of a legal system that actually generated moral obligations (Finnis 2011: 3, 15). Finnis even claims that Hart and Raz are *sotto voce* practitioners of this same method. For Hart and Raz, significance and importance are to be assessed from the perspective of those who take the "internal point of view" towards their society's law, meaning those who treat the law as a genuine standard of conduct, use it to guide their conduct, and appeal to it to evaluate the conduct of their peers (Hart 1994: 89–90). One of Hart's lasting contributions to legal philosophy is his observation that we cannot understand law if we fail to account for the way legal norms are used as reasons for action and normative standards "in the lives of those who normally are the majority of society" (Hart 1994: 90). For this reason, the central case of law (as Finnis reads Hart) cannot be a legal system in which the majority of society sees the law merely as a system of state-issued threats or predictions about when coercive force will be used against them. Rather, the central case of a legal system is one whose norms are generally treated as reasons for doing or refraining from certain acts, and reasons for criticizing or punishing non-conforming conduct.

Hart and Raz both noticed, though, that there are many motivations for treating law this way. One might take oneself to have a general moral obligation to obey the law, and view the law as a standard of conduct for that reason, but one could also have less lofty motives. For example, one could take the internal point of view out of unreflective habit, or out of a desire to get along with one's peers, or out of a sense of professional duty or consideration of long-term self-interest. But for the purposes of legal theory, these different motivations are explanatorily idle. Differentiating between them does not help us explain any interesting *general* feature of legal systems.

Finnis thinks that this is where Hart and Raz go wrong. There are central and peripheral cases *of the internal point of view itself*, and we have to discriminate between them to identify the central case of law. Hart and Raz, on his view, stop arbitrarily with those who merely treat the law as creating obligations, whereas the central case of the internal point of view for Finnis is the perspective of one who *correctly* treats the law as morally obligatory (which we will call *the moral point of view* on law). Although it is possible to treat the law as a standard of conduct out of habit or long-term self-interest, these are, for Finnis, "deviant" cases of recognizing something as a genuine reason for action. That is, these attitudes towards the law can only be understood as defective, corrupted versions of the moral point of view. Hart and Raz correctly see that the central case of law cannot be, as O. W. Holmes arguably thought, the case of someone who obeys the law only when afraid of sanction, but they then incorrectly treat as central the case of someone who simply treats the law as obligatory, even when it is not really morally obligatory.[12]

Why do we need to privilege the moral point of view to identify the central case of law? According to Finnis, this point of view is the one "that brings law into being as a significantly differentiated type of social order and maintains it as such" (Finnis 2011: 14). Other points of view "will, up to a point, maintain in existence a legal system … if one already exists. But they will not bring about the transition from the pre-legal (or post-legal!) order" (Finnis 2011: 14). If this is an empirical claim about how legal systems emerged in human history, it is both implausible and unsupported. More importantly, it is also of unclear relevance to the question what the concept of law in modern societies is. But perhaps Finnis does not mean it as an empirical claim: he also says that it is only from the moral point of view that it is "a matter of *overriding importance* that law as distinct from other forms of social order should come into being, and thus become an object of the theorist's description" (Finnis 2011: 15, emphasis added). The suggestion, it seems, is that we can only assess the good reasons for having law from the moral point of view. That might be true, but it is, again, unclear why we need to know the good reasons for having law in order to assess which of its features are important for the legal theorist. There appears to be no answer to this question that does not beg the question against methodological positivism. The answer cannot be that we cannot understand law without understanding its moral value, for that is precisely what this argument is supposed to show. It is hard to resist the conclusion that Finnis has simply changed the subject: what began as an argument about how to justify judgments of salience has now become an argument about how to assess the moral value of law, with no motivation for linking the two considerations (see Leiter 2007: 167–8 for a related objection).

Liam Murphy presents a different kind of objection to descriptive jurisprudence (Murphy 2001). According to Murphy, although there is widespread agreement about many aspects of law, there is also intractable disagreement about many of the most important questions of legal theory. Murphy cites the pervasive disagreement among legal theorists over whether the laws of any given system can be determined by moral criteria as

[12] We are merely trying to describe Finnis's argument here, though it is clear it misdescribes Hart's argument. Hart's argument is not that the point of view of those who treat the law as obligatory is morally superior to the point of view of the Holmesian "bad man," it is that it is descriptively false that everyone, legal officials included, adopts the point of view of the "bad man."

opposed to facts about the behavior and attitudes of officials as evidence of the thinness of our folk concept (Murphy 2001: 381). The existence of intractable disagreement about such issues implies that the folk concept will be silent about them, and so we cannot hope to answer these questions by analyzing our shared concept of law. Therefore, a successful theory of law cannot be merely a descriptive theory of the folk concept of law, because such a theory will be rather obviously incomplete.

Murphy's proposed solution is to abandon methodological positivism in favor of what he calls a *practical-political methodology*, whereby questions about the nature of law are decided according to which answers will yield the best moral or political outcomes. On Murphy's view the legal theorist should ask for a significant range of questions which of any number of competing theories of the nature of law, if adopted, would yield the best moral or political consequences? Thus, for example, Murphy endorses Joseph Raz's thesis that all legal norms are valid solely in virtue of their social pedigree, over Dworkin's claim that a norm is legally valid if and only if it follows from the best constructive interpretation of a society's legal practices, on the grounds that the former will have more desirable political consequences (Murphy 2001: 408).

The problems with this kind of methodological position are obvious and well-documented (see Soper 1986; Dickson 2004: 147–9). A theory of law is a theory of what we should *believe* are the central features of a particular kind of social institution, but Murphy's practical-political methodology suggests that it is up to the legal theorist to decide what to believe—but belief does not work that way! The truth of a theory of law is not determined by the moral or political consequences of widespread adoption of that theory. To suppose otherwise turns large swaths of legal theory into wishful thinking.

Finally, we turn to Ronald Dworkin's challenge to methodological positivism. He rejects methodological positivism on the grounds that the concept of law is an *interpretive concept*. On Dworkin's view, it is a crucial feature of the folk concept of law that it includes what he calls an "interpretive attitude." Participants in a social institution evince an interpretive attitude towards that institution just in case (1) they see it as having some value or as serving some purpose, and (2) they take the requirements of that institution to be partly determined or constrained by what would count as fulfilling that purpose (Dworkin 1986: 47). Law satisfies the first condition because the participants in a legal system generally take their law to serve the purpose of justifying state coercion (Dworkin 1986: 98, 109). In Dworkin's words, "[l]aw insists that force not be used or withheld … except as licensed or required by individual rights and responsibilities flowing from past political decisions about when collective force is justified" (Dworkin 1986: 93). (All positivists deny this, and Dworkin never offers any satisfactory explanation for what is essentially a stipulation on his part.) Law satisfies the second condition because the participants take this scheme to be partly determined by the principles of political morality that do in fact justify the use of coercion in light of past political decisions. More concretely, it is Dworkin's view that a society's law is the scheme of rights and duties that figure in or follow from the explanation that best fits and best justifies that society's institutional history (i.e. its legislative enactments, judicial opinions, constitution, etc.).[13] This style of explanation is what Dworkin calls *constructive interpretation*.

[13] It justifies the institutional history with respect to the purpose of justifying state coercion.

Dworkin believes that law's interpretive nature has profound implications for the methodology of legal philosophy. In particular, all "useful" theories of law for Dworkin must themselves be constructive interpretations of some particular society's institutional history (Dworkin 1986: 102, 108–10). The legal theorist's task on this view is to offer an account of what the law of his or her community is, or a theory of how disputes about the law are to be decided in that community. "Jurisprudence," Dworkin says, "is the general part of adjudication, silent prologue to any decision at law" (Dworkin 1986: 90). Positivism may be useful for identifying the institutional history of a legal system, but it does not tell us what the law really is, since that is what follows from constructive interpretation.

This methodological stance makes it difficult to put Dworkin's views into conversation with other theories of law. Debates in legal philosophy are almost exclusively debates about how to formulate a general theory of law, and substantive debates are almost exclusively about the general features of legal systems. By building in "justifies coercion" to the concept of law, most legal theorists think Dworkin has limited his theory to a subset of legal systems in which legal coercion is morally justified, thus making it a case of "particular jurisprudence" (see Hart 1994: 240–2; Leiter 2007: 159). Dworkin sometimes suggests that all viable theories of law must be constructive interpretations because any other kind of theory is doomed to fail. Many theorists think this is a *reductio* of Dworkin's view since it entails that everyone in a legal system could, in fact, be mistaken about *what the law is*, since they might not realize what the best constructive interpretation would say the law really is![14]

Dworkin's strongest argument for understanding law as an interpretive concept claims that it allows us to make better overall sense of the actual practices of judges, lawyers, and other legal professionals.[15] In particular, it does a better job of explaining the existence and

[14] "Soft" positivism, which allows that the rule of recognition could incorporate moral criteria of legal validity, actually faces the same problem, one of several reasons for thinking that Hart was mistaken in his posthumous Postscript in thinking that was a viable response to Dworkin's early criticisms.

[15] Dworkin also offered the argument known as the "the semantic sting." In *Law's Empire*, Dworkin argues that all non-interpretive theories of law must be theories of the semantics of legal terms (Dworkin 1986: 31–44). He furthermore takes all such theories to be committed to a *criterial semantics*: the thesis that an analysis of the meanings of legal terms is an analysis of the shared criteria used by lawyers, judges, and other legal professionals to apply these terms (Dworkin 1986: 35–6). But this combination of views is untenable, because it cannot adequately explain the fact that legal professionals often disagree about what the law is. "Semantic theories", as Dworkin calls them, are committed to saying that when professionals disagree about whether some proposition of law is true (e.g. "it is illegal to drive over 45 mph on Lakeshore Drive"), they must be using different criteria to apply the relevant legal terms. But then it follows that they must *mean different things* by their terms, and that they therefore are not really disagreeing, which is absurd (Dworkin 1986: 43–5). All non-interpretive theories fall prey to this semantic sting, and the only plausible alternative to a semantic theory of "law" is an interpretive theory.

The semantic sting fails in myriad ways, as several commentators have observed (see Raz 2001, Coleman and Simchen 2003, and Shapiro 2007 for a thorough discussion). Most immediately, there is no reason to think that all non-interpretive theories of law are committed to criterial semantics, as they must be if they are to be stung by Dworkin's argument. More generally, there is no reason why all non-interpretive theories must be theories of the semantics of legal terms. Indeed, Hart specifically disavows any real concern with the norms governing the application of legal terms in the opening sections of *The Concept of Law*. Dworkin must be basing this conclusion on Hart and others'

nature of *theoretical disagreements* about law. Disagreement about the law is, of course, a familiar fact, but Dworkin observes that these disputes can come in two forms. A *theoretical disagreement* is a disagreement about the criteria of legal validity of a particular legal system, that is, about what makes a particular proposition of law true in a given society. For example, if we disagree about whether the fact that the city council's vote on the speed limit suffices to set that limit, we are having a theoretical disagreement. This type of disagreement contrasts with what Dworkin calls *empirical disagreement* about the law, in which parties agree about the criteria of legal validity but disagree about whether those conditions set out by those criteria have been satisfied. Thus if we agree that a city council vote would suffice to make the speed limit on Lakeshore Drive 45 mph, but disagree about whether such a vote has actually taken place, we are having an empirical disagreement.

Dworkin argues that theoretical disagreement is a significant feature of law, citing, in particular, disagreement about the canons of statutory interpretation among (American) appellate court judges as evidence. Moreover, positivists appear to have a difficult time accounting for this type of disagreement. If positivism is correct, then whatever counts as law in a particular society is determined exclusively by facts about official behavior and attitudes, the social facts that constitute the rule of recognition. If officials do not, in fact, converge on the criteria of legal validity, then there is no fact of the matter about what the criteria of legal validity are, and therefore no fact about which to have a genuine disagreement. Positivists therefore appear to be committed to saying that in all cases of theoretical disagreement, the parties to the disagreement are either mistaken or disingenuous (Leiter 2009: 1224–6; Shapiro 2011: 290–1). And if theoretical disagreement is as common as Dworkin suggests, it follows on the positivist view that a typical legal system is characterized by massive amounts of error or disingenuity about what the law is.

Dworkin takes this to be an unpalatable consequence of legal positivism and its methodological posture. On the surface, judges appear to disagree about what the law requires because they disagree about the criteria of legal validity. And in resolving such disputes, they at least claim to be deciding according to what the criteria of validity *really are*. Methodological positivism demands a debunking explanation of what is going on in such cases, and this, Dworkin thinks, counts against positivism and in favor of his theory, which offers, he thinks, an explanation of theoretical disagreement more faithful to a literal, rather than a debunking, interpretation of the disagreement. When judges disagree about the criteria of legal validity, they are disagreeing about what counts as the best constructive interpretation of their society's institutional history, and the existence of disagreement does not rule out the possibility of a right answer to this question. Understanding law as an interpretive concept, then, allows us to preserve the surface features of legal argument and disagreement.

Positivism is, sometimes, committed to explaining away theoretical disagreement,[16] though the theoretical costs of this commitment are proportional to how common that

endorsement of linguistic analysis as an important tool of legal theory. However, the usefulness of linguistic analysis, as we saw earlier, is not limited to its ability to elucidate the meanings of terms.

[16] Some legal systems, in fact, have pedigreed interpretive rules for resolving such disagreements: e.g. in Canadian constitutional law, an argument that appeals to the original intent of the drafters of the Charter is forbidden. There can be no theoretical disagreement about this kind of interpretive issue in Canada, even though it is rife in American constitutional law.

disagreement really is. If it turns out to be a relatively marginal feature of law, the costs of adopting a debunking explanation might be acceptable in light of positivism's other theoretical virtues. Dworkin's case for the pervasiveness of theoretical disagreement rests on his analysis of several Anglo-American legal cases, though questions have been raised about the adequacy of his reading of those (Leiter 2009: 1232–48). More importantly, appellate cases represent only a tiny fraction of all litigated disputes, and even a smaller fraction of all legal disputes (Leiter 2009: 1226–7). Furthermore, appellate courts choose to hear these cases precisely because of their novelty relative to typical legal disputes. They tend to be cases that stand out because they point to areas where the law is particularly unclear or unsettled. By contrast, the vast majority of cases never make it to trial or never make it onto the docket of a court of appeals precisely because there is no real question about what the law requires in these cases. The most obvious explanation for this is that the parties to a dispute *agree about what the law is*. Functioning legal systems are most notable for the massive amounts of agreement about the law, which is exactly what a positivist theory predicts (Leiter 2009: 1227). So, contra Dworkin (and others sympathetic like Shapiro 2007, 2011), theoretical disagreement is not nearly so pervasive, and where it does occur, the debunking explanations are actually the most plausible.

All of the objections to methodological positivism considered in this section are alike in that they each rely on premises about the existence of some form of disagreement, and that they each rely on premises that cannot be confirmed from the armchair. They are all motivated by some putative observations about social facts, observations for which there can be no a priori argument. We have offered some preliminary reasons for thinking that methodological positivism can survive these challenges, but ultimately these disputes can only be decided by closer attention to the social facts that all parties agree partly constitute a legal system.

4. Conclusion

We have focused in this essay only on the methodology for the central topics in philosophy of law: the nature of law and the relationship between legal and other norms, especially moral ones. But much work in philosophy of law concerns philosophical questions about substantive legal doctrines, questions like: Are contracts morally binding promises, or agreements that can be breached to yield efficient outcomes? What is the difference between excuse and justification in the criminal law? Is the law governing private wrongdoing (the law of "torts") implementing a kind of corrective justice, or is it better explained in terms of economic efficiency? This work has a great advantage over the work in the core of legal philosophy, because it has an undisputed datum: a substantive body of doctrine enacted by legislatures and courts in various jurisdictions. The ambition of such work is almost always the same: to try to identify the moral (or at least normative) coherence of the substantive doctrine. It also almost always confronts the same problem: the law is morally incoherent, since it has been created over a long period of time, by persons with many different interests and concerns. Thus this work typically shifts to an explicitly normative

perspective that raises the usual methodological issues in normative philosophy, but these are beyond the scope of this essay.[17]

Bibliography

Coleman, Jules and Ori Simchen. 2003. "'Law,'" *Legal Theory* 9: 1–43.
Dickson, Julie. 2001. *Evaluation in Legal Theory*. Oxford: Hart Publishing.
Dickson, Julie. 2004. "Methodology in Jurisprudence: A Critical Survey," *Legal Theory* 10: 117–56.
Dworkin, Ronald. 1986. *Law's Empire*. Cambridge, MA: Harvard University Press.
Finnis, John. 2003. "Law and What I Truly Should Decide," *The American Journal of Jurisprudence* 48: 107–29.
Finnis, John. 2011. *Natural Law and Natural Rights*. 2nd edition. Oxford: Oxford University Press.
Green, Leslie. 1996. "The Concept of Law Revisited," *Michigan Law Review* 94: 1687–717.
Green, Leslie. 2010. "Law as a Means." In *The Hart-Fuller Debate in the Twenty-First Century*. Edited by Peter Cane. Oxford: Hart Publishing: 169–88.
Hart, H. L. A. 1994. *The Concept of Law*. 2nd edition. Edited by Joseph Raz and Penelope Bullock. Oxford: Clarendon Press.
Hart, H. L. A. 2001. *Essays in Jurisprudence and Philosophy*. Oxford: Oxford University Press.
Jackson, Frank. 1998. *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Oxford University Press.
Kelsen, Hans. 1967. *Pure Theory of Law*. Translated by Max Knight. Berkeley: University of California Press.
Leiter, Brian. 2007. *Naturalizing Jurisprudence*. Oxford: Oxford University Press.
Leiter, Brian. 2009. "Explaining Theoretical Disagreement," *The University of Chicago Law Review* 76: 1215–50.
Marmor, Andrei. 2009. *Social Conventions: From Language to Law*. Princeton: Princeton University Press.
Murphy, Liam. 2001. "The Political Question of *The Concept of Law*." In *Hart's Postcript: Essays on the Postcript to* The Concept of Law. Edited by Jules Coleman. Oxford: Oxford University Press: 371–409.
Murphy, Mark. 2006. *Natural Law in Jurisprudence and Politics*. Cambridge: Cambridge University Press.
Perry, Stephen. 2001. "Hart's Methodological Positivism." In *Hart's Postcript: Essays on the Postcript to* The Concept of Law. Edited by Jules Coleman. Oxford: Oxford University Press: 311–54.
Raz, Joseph. 2001. "Two Views of the Nature of the Theory of Law: A Partial Comparison." In *Hart's Postcript: Essays on the Postcript to* The Concept of Law. Edited by Jules Coleman. Oxford: Oxford University Press: 1–37.

[17] We are grateful to Thomas Adams, Leslie Green and Mark Murphy for helpful comments on an earlier draft; to participants in a Law & Philosophy Workshop at the University of Pennsylvania for comments and questions on a later draft, to Matt Lister for written comments on that same version, and to Tamar Gendler for advice about the penultimate draft.

Raz, Joseph. 2009a. "Can There be a Theory of Law?" In Raz, *Between Authority and Interpretation*. Oxford: Oxford University Press: 17–46.

Searle, John R. 1995. *The Construction of Social Reality.* New York: The Free Press.

Searle, John R. 2009. *Making the Social World: The Structure of Human Civilization*. Oxford: Oxford University Press.

Shapiro, Scott J. 2007. "The 'Hart-Dworkin' Debate: A Short Guide for the Perplexed." In *Ronald Dworkin*. Edited by Arthur Ripstein. Cambridge: Cambridge University Press.

Shapiro, Scott J. 2011. *Legality.* Cambridge, MA: Harvard University Press.

Soper, Philip. 1986. "Choosing a Legal Theory on Moral Grounds," *Social Philosophy & Policy* 4: 31–48.

Weinberg, Jonathan, Shaun Nichols, and Stephen Stich. 2001. "Normativity and Epistemic Intuitions," *Philosophical Topics* 29: 429–60.

CHAPTER 35

FEMINISM

ISHANI MAITRA

1. Introduction

Feminist philosophy, as I understand it, is not a distinct branch of philosophy, akin to ethics or epistemology. Rather, feminist work in philosophy seeks to bring to the fore a set of questions, about the role of gender in our various practices, and offers a set of tools with which to approach those questions.[1]

To see what I have in mind, let's take feminist work in epistemology as an example.[2] If we understand epistemology as the study of our epistemic practices—what knowledge is, who knows and under what circumstances, how we come to know when we do, and so on[3]—then feminist epistemology can be seen to be concerned with the following pair of big-picture questions:

> *Descriptive question*: What is the role of gender in our epistemic practices?
>
> *Normative question*: What should the role of gender be in our epistemic practices?

These questions are concerned with gender in all of its aspects; that includes gender norms, gender identity, gendered social roles, and many others (Haslanger 2000). By comparing what is said in response to each question, feminist epistemologists seek to uncover the ways in which our epistemic practices systematically and unjustly disadvantage some on

[1] As I'll note later in this section, this is not the only way to view the project of feminist philosophy, but I do think it's a useful one. Others have suggested similar views. For example, in their introduction to the *Cambridge Companion to Feminism in Philosophy*, Miranda Fricker and Jennifer Hornsby explain that they "wish neither to advocate an understanding of feminist philosophy as a separate and distinctive branch of philosophy, nor to argue for the ability of feminist philosophy to replace philosophy" (Fricker and Hornsby 2000, 1).

[2] My discussion of feminist epistemology—and of feminist work in philosophy more generally—owes a great deal to Elizabeth Anderson's excellent survey in Anderson (2012).

[3] When talking about feminist epistemology in this chapter, I'll largely focus on questions about knowledge. But this is a simplification: feminist epistemologists are also interested in questions about other central epistemic concepts like belief, justification, reasons, rationality, objectivity, and the rest.

the basis of gender, while systemically and unjustly privileging others on the same basis (Anderson 2012, Rooney 2012).

Not all disadvantage is unjust, of course; neither is all privilege. So part of the project here is to discriminate the just advantage (or disadvantage) from the unjust, in a principled way. And feminist philosophers have disagreed about which advantages are just, and which unjust. But insofar as feminist epistemology is concerned with matters of justice in this sense, it is an unabashedly political project.

Both the descriptive and the normative questions have been crucially important in feminist epistemology. On the descriptive side, feminist epistemologists have developed tools to help identify how gender is implicated in both (i) who is taken to know, that is, who is considered a *subject* of knowledge, and (ii) who (and what) we have knowledge about, that is, who (and what) constitute the *objects* of our knowledge.

There is any number of examples of this kind of work, both recent and not-so-recent. But I'll choose a few recent ones to illustrate the point: Nancy Tuana has discussed how the women's health movement of the 1970s and 1980s fought to render women as subjects—and not merely objects—of knowledge about their own bodies and health (Tuana 2006). Londa Schiebinger has argued that gender dynamics and other power relations in colonial Europe prevented the transfer of significant medical knowledge from the European colonies to the colonial powers (Schiebinger 2004). Miranda Fricker has described how sexist stereotypes contribute to women testifiers being unjustly afforded less credibility than the evidence demands (Fricker 2007). And Jose Medina has highlighted how sexism (and racism) in the "social imaginary" can get in the way of both knowledge of others as well as self-knowledge (Medina 2013).

As might be clear from even this brief review, feminist work in epistemology draws heavily on work in other disciplines, including history, legal theory, medicine, political science, and sociology, to name a few. That should not be surprising, since each of these disciplines is concerned with some aspects of our epistemic practices, and so, can help shed light on the role gender plays in those practices. This kind of inter-disciplinarity is a characteristic of much feminist work in philosophy (Richardson 2010).

Moving now to the normative question, feminist epistemologists have also offered prescriptions for where gender should, and should not, be involved in our epistemic practices. In much of the recent work already mentioned, it is argued that gender operates to unjustly epistemically disadvantage many women, and some men; therefore, there is reason to limit its impact in those areas. But theorists have differed on how that should be achieved.

Consider, for example, the problem of ensuring that sexist stereotypes don't unduly interfere with the credibility afforded to women. In response, Fricker has argued that we *qua* individual recipients of testimony should cultivate certain epistemic virtues, in particular, the virtue of testimonial justice (Fricker 2007). Karen Jones has offered a set of guidelines that can be put to use in determining credibility (Jones 2002). Medina has highlighted the prejudice-busting potential of the kind of friction that comes from interacting with different ways of viewing the world (Medina 2013). And Linda Alcoff has argued that we must go beyond individual remedies to structural ones (Alcoff 2010).

The work I've discussed so far has mostly focused on how women are systematically disadvantaged by the operations of gender in our epistemic practices. But some feminist epistemologists have argued that gender can also be a source of epistemic privilege for

women. Most famously, feminist standpoint theorists have defended the view that gender can be partly constitutive of (epistemically) privileged standpoints on some subject matters (Hartsock 1987, Collins 1990, Harding 1991, among others). For example, some standpoint theorists have argued that women's status as targets of sexist (and other) oppression renders them better able to grasp the true nature of social reality than those who stand to benefit from that oppression. Later versions of feminist standpoint theory have developed these insights in new directions, including by emphasizing the contingent, that is, non-automatic, nature of the privilege conferred by a given standpoint. (See Wylie 2003 for a very useful discussion of this and other developments in standpoint theory.)

Summing up this part of the discussion: on the conception I've been outlining, feminist epistemology is not distinguished by focus on any particular topic within epistemology, or by commitment to any particular view about, say, what knowledge consists in. Nor are feminist epistemologists committed to there being anything distinctive about what women (or men) know, or about how they come to know it. (See Rooney 2012 for an interesting discussion of misconceptions about feminist epistemology.) Rather, as my examples of recent work show, feminist epistemologists address a wide range of topics within epistemology. They do this with an eye to uncovering systematic and unjust privilege/disadvantage on the basis of gender. As such, feminist epistemology is a deeply value-laden inquiry, but not particularly mysterious for that reason.

I took epistemology as my initial example because feminist work in that area is particularly well developed. But much feminist work in other areas of philosophy can be understood under the same rubric. For example, feminist philosophy of language can be understood as addressing the descriptive and normative questions with respect to our linguistic practices, while feminist work in ethics can be seen as addressing their analogues with respect to our ethical practices. These areas might differ, however, with respect to whether the emphasis in a given area is on the descriptive question or on the normative one.[4]

Understanding the role of gender in our various practices requires that our tools—our concepts, frameworks, and theories—be adequate to the job. Therefore, a central concern of feminist philosophy has been whether existing philosophical tools reveal, rather than obscure, the operations of gender in our practices. To put it another way, feminist philosophy seeks to uncover and theorize sexism in existing philosophical work, and to correct for it where possible.[5] As such, it is—to borrow some terminology from Louise Antony—a *critical* project (Antony 2012).

[4] I note that there will be some strain in fitting all feminist philosophical work into the rubric I've been describing. Consider, in particular, feminist metaphysics, which has been concerned with understanding categories central to feminist theorizing, including gender, but also social kind, social structure, social construct, and so on (Haslanger 2012, the essays collected in Witt 2011a, Witt 2011b, Barnes 2014, among others). Some of this work can be understood to be examining the role of gender in our metaphysical practices, where those include our practices of classifying individuals as persons or not, as men, women, or neither, as disabled or not, and so on. I don't have the space here to consider whether all work in this area can be so captured.

[5] I take 'sexism' here to mean, roughly, unjust and systematic disadvantage on the basis of gender. Sexism thus understood needn't be intentional, though of course, it can be. See §3 for more on sexism, or sexist oppression.

Antony cites Sandra Harding and Merrill Hintikka's classic anthology, *Discovering Reality*, as a founding document for this sort of philosophical work. Many of the essays in Antony and Charlotte Witt's important collection, *A Mind of One's Own*, can be described this way as well. But as before, we can also illustrate the critical aspects of feminist philosophy from more recent work: Cynthia Townley has criticized epistemologists' exclusive focus on knowledge at the expense of ignorance; drawing on earlier work by Lorraine Code, Townley argues that we need to properly value ignorance in order to develop more responsible, that is, non-oppressive, epistemic practices (Townley 2006, 2011, Code 1987, 1991). Kristie Dotson has raised concerns about the use of "closed conceptual structures", roughly, accounts that purport to be complete with respect to some given phenomena when they are in fact partial; on Dotson's view, the use of these structures can contribute to epistemic oppression (Dotson 2012). And Elizabeth Anderson has criticized philosophers of science, feminist and otherwise, for failing to explain why it might be better, that is, more epistemically fruitful, to rely on non-sexist values rather than sexist ones in choosing background assumptions for scientific inquiry (Anderson 2004).

Feminist philosophers have also engaged in critical projects with respect to other feminist work, both within and outside philosophy. Some of the work described in the previous paragraph qualifies on that count as well. So does Mariana Ortega's argument that white feminists, while recognizing differences among women in the abstract, have still failed to come to terms with their lack of knowledge about the particular experiences of women of color; Ortega distinguishes between arrogant perception and loving, knowing ignorance, and exhorts white feminists to practice more of the latter (Ortega 2006).

In explaining this notion of a critical project, Antony distinguishes between that kind of project and a *replacement* one (Antony 2012). A replacement project, as the label suggests, aims to replace existing philosophical work from the ground up with a feminist—or more broadly non-oppressive—version. Such a project might be motivated by the conclusion, arrived at as the result of a prior critical project, that the relevant work is deeply and ineliminably sexist, and the only plausible remedy is to start over from self-consciously feminist premises.

Either kind of project is potentially transformative, in the sense that either might end up requiring large-scale changes in our philosophical practices. But they differ at least with respect to their starting points: the critical project begins by trying to eliminate sexism where possible, while the replacement project takes such a bit-by-bit approach to be impracticable, or worse. In my view, quite a lot of feminist work in philosophy—including many of the views that I'll be discussing in the rest of this chapter—can be characterized as engaging in the critical project.[6]

[6] Antony also distinguishes a third kind of project, namely, a *practical* project, alongside the critical and replacement ones mentioned above (Antony 2012). A practical project is one which applies existing philosophical tools to topics that have some special urgency for women. Antony mentions Judith Jarvis Thomson's "A defense of abortion" and Ann Ferguson's "On conceiving motherhood and sexuality" as exemplars of this type of work. Antony is surely right to claim that much feminist philosophy is engaged in such practical projects, though (as she would no doubt agree) not all practical projects are in any interesting sense *feminist*: after all, they can be undertaken with the aim (and/or effect) of perpetuating sexism. I hope that something like the rubric I've suggested near the beginning of this chapter can help characterize what makes some practical projects feminist, though again, I don't have the space to pursue that point here.

Success in the critical project requires engaging with existing philosophical work while avoiding what Ann Cudd has described as "the twin dangers of accommodation and co-optation" (Cudd 2012, 18). Accommodation in this sense involves "adapting one's own perspective to concur with that of the dominant", and thereby failing to identify sexism and other forms of oppression (2012, 18). Co-optation involves "taking on the projects of the dominant group as one's own", and thereby not only failing to see sexism, but actually helping to perpetuate it (2012, 18). Accordingly, feminist philosophers engaging in critical projects must retain something of the perspective of the outsider even while engaged with familiar philosophical questions.

Thus far, I've been focusing on some *questions* that are central to feminist work in philosophy. But besides drawing attention to these questions, feminist philosophers (and feminist theorists more generally) have also developed a distinctive set of tools for addressing them. Some of these tools—for example, the sex–gender distinction—largely originated in the work of feminist theorists; others—for example, objectification, autonomy—did not, but have been developed by feminist philosophers in distinctive directions.

In what follows, I'll discuss some of these tools, try to indicate *why* they've been important, and look at some recent debates on how they're best understood. The ones I'll focus on are: the sex–gender distinction; oppression; objectification; and situated knowledge (and situated agency). Besides these, there are others that are also important to feminist theorizing—social construction, social power, silencing, embodiment, ideology, care, and autonomy, to mention just a few more—that can't be included here for reasons of space. But I take the ones I focus on to be specially important to the feminist project, in ways that will emerge in my discussion.

Much of what I say here will be familiar to those working in feminist philosophy. But I hope that the discussion will nonetheless be valuable by identifying some methodological inclinations in feminist philosophy, in the form of a number of themes that recur through work in the area. I'll finish the chapter by describing some of those recurring themes.

A final preliminary point, before proceeding: my focus in this chapter will be on feminist work in philosophy in the *analytic* tradition, very broadly construed. This is also largely for reasons of space: given the great range of topics and approaches exemplified in feminist work, some restrictions of scope are necessary. But that means that the picture presented in this chapter is crucially partial, leaving out some very significant feminist philosophical work.

2. Sex and Gender

The distinction between sex and gender has a long history in feminist work, both within and outside philosophy. In fact, it has been suggested that

> in a sense, feminism *begins* with the distinction between sex as a biological category and gender, the hierarchical distinction between men and women, masculine and feminine, as socially constructed.
>
> (Cudd 2012, 19, emphasis added)

As Cudd suggests in the passage above, this distinction originally arose as one between a purely biological category, and a social one. Sex (male, female) was understood to be determined by biological features, including reproductive organs, other anatomical features, chromosomes, and the like. By contrast, gender (man, woman, perhaps others) was taken to be deeply social, though still importantly related to the biological. These dual aspects of gender are highlighted by the slogan "gender is the social meaning of sex".

In using the terms 'sex' and 'gender' in this way, feminist theorists arguably depart from ordinary usage, where these terms are often used interchangeably (Saul 2006). A big part of the point of drawing this distinction is to resist what is sometimes called 'biological determinism', roughly, the view that biological differences between men and women determine—and *should* determine—their different social standings (Mikkola 2012). By insisting on the distinction between the biological and the social, and by pointing to differences in how sex is interpreted in different socio-cultural contexts, feminist theorists have argued against this kind of determinism.

Theorists who agree that gender is partly constituted by our social and cultural practices—that is, that gender is *socially constructed*—have nevertheless differed on what that construction consists in. Accordingly, a broad range of views have been offered about precisely *how* our social and cultural practices constitute gender, and so, about how gender relates to sex. (See Mikkola 2012 for a very helpful survey.) These theories differ in a number of ways, including with respect to which aspect(s) of gender they take to be explanatorily basic. But they have typically argued that gender is not just a social, but also a *hierarchical* category: men and women don't just occupy different social roles or social positions, but ones that are unequal in terms of the power and privilege accruing to them.

More recently, the sex–gender distinction has been criticized, on both sides. On the construal of sex, it has been argued that our social practices are crucially implicated in who counts as male and who counts as female, just as they are in who counts as belonging to which gender. Influenced in part by biologist Anne Fausto-Sterling's work on how the 'two-sex system' is maintained, some theorists have concluded that sex is just as deeply social as gender, and therefore, that the original understanding of the sex–gender distinction cannot stand (Fausto-Sterling 1993, 2000).

The category of gender has also come under scrutiny. One major source of concern has to do with whether any attempt to give an account of how each gender is constituted will inevitably end up excluding some who ought not be excluded (Spelman 1988). As Jennifer Saul puts it,

> Once we reflect on the wide variety of social roles lived out by women (consider a Somali refugee mother of five, the Queen of England, and a childless Californian lesbian artist, for a few examples), it becomes clear that there is no one 'women's social role'.
>
> (Saul 2012, 197)

Given that the category 'woman' encompasses historical eras, races, sexualities, classes, disabilities, and nationalities, we might worry that no substantive account of the category can avoid this problem of exclusion.

In response to these and related worries, some have concluded that feminists should give up on the sex–gender distinction altogether (Moi 1999, chapter 1, Mikkola 2011). But of course, doing so has serious downsides; among them is the worry that without some

understanding of gender—and especially the category 'woman'—the goals of feminism may prove impossible to articulate (Alcoff 2006, chapters 5 and 6).

Some recent accounts have focused particularly on developing understandings of gender that can (i) address the problem of exclusion just described, and (ii) serve some, but perhaps not all, of the theoretical and practical purposes we have in employing gender categories. I'll end this section by considering two such theories, both of which aim to offer accounts of gender that can serve as a basis for feminist theorizing (and political action).

In her recent book, Witt has offered an account that is intended to make sense of the thought, shared by many of us, that gender is partly constitutive of who we are (Witt 2011b). Witt's broadly Aristotelian view articulates a kind of gender essentialism—which she labels 'uniessentialism'—to capture this thought.

Witt distinguishes between human beings, persons, and social individuals, three different kinds of entities that can nonetheless share a single spatio-temporal location. The question of gender essentialism arises for social individuals, not the other kinds of entities.

Qua social individual, each of us occupies multiple social roles. Thus, for example, I might occupy the social roles of professor, employee, spouse, parent, child, and so on. Each of these roles brings with it a set of norms. But, Witt argues, something must organize this bundle of social roles (or sets of social norms) into a single individual, to whom all the norms apply, and for whom they can come into conflict. For example, the norms pertaining to the spouse role might sometimes conflict with the norms pertaining to the employee role. Gender is the function—a distinct function for each gender—that organizes and unifies all our social roles into the particular social individuals we are. It thus creates and maintains the 'normative unity' that is the social individual. In addition, it also 'inflects' the other social roles: which norms pertain to me in my role as professor depends in part on my gender. In light of these aspects, gender is not only a social role, but a *mega* social role, and also uniessential to us *qua* social individuals. (See Sveinsdóttir 2012 for a particularly lucid discussion of Witt's project, and some important criticisms.)

In a series of papers, Sally Haslanger has given what she describes as a 'focal analysis' of gender as a social class; other aspects of gender, such as gender identity and gender norms, are to then be understood by reference to this focal analysis (Haslanger 2000). For Haslanger, gender categories pick out positions within a social hierarchy. More specifically, they pick out social positions based on prevalent assumptions about the meanings of sex differences:

> S *is a woman* iff$_{df}$ S is systematically subordinated along some dimension (economic, political, legal, social, etc.) and S is 'marked' as a target for this treatment by observed or imagined bodily features presumed to be evidence of a female's biological role in reproduction.
>
> S *is a man* iff$_{df}$ S is systematically privileged along some dimension (economic, political, legal, social, etc.) and S is 'marked' as a target for this treatment by observed or imagined bodily features presumed to be evidence of a male's biological role in reproduction.
>
> (Haslanger 2003–4, 6–7)

Responding to the worry about exclusion mentioned earlier, Haslanger notes that her definition is intended to serve a specific purpose, namely, helping us understand (and end) sexist oppression; she argues that it excludes from the category 'woman' only those who are

not oppressed *qua* women. (See Saul (2006) and Glasgow (2009) for some trenchant criticisms of Haslanger's account of gender and her parallel account of race.)

As we can see, there are important differences between Witt's account of gender, and Haslanger's. But there are some significant similarities as well. One similarity is that both views address the problem of exclusion by offering highly abstract accounts of gender. A second similarity is the emphasis on social structures. On both views, what it is to be a woman (or a man, or any other gender) depends on available social meanings or social roles, not merely on one's own preferences and inclinations. On both views, feminist political action will require changing existing oppressive social structures.

3. Oppression

Feminism, bell hooks has famously said, is "the struggle to end sexist oppression" (hooks 2000, 28). She notes that feminism is more popularly characterized as a movement to secure women's equality to men, but argues that we should reject this alternate characterization, for several reasons. One major reason has to do with the fact that men themselves are not equal; some men are more privileged than others due to their race, class, sexuality, and so on. But if 'equality' here is to mean equality to men of one's own race, class, sexuality, etc., then feminism will be mostly attractive to the most privileged women.

This characterization of feminism has been very influential. But for it to be able to do interesting philosophical work, it requires an understanding of what oppression consists in, and how we might distinguish sexism (or sexist oppression) from racism, homophobia, transphobia, classism, and so on.

On one way of thinking about oppression, an individual (or group of individuals) oppresses another individual (or group of individuals). Following Haslanger, we might call this 'agent oppression', and render it as follows:

> *x* oppresses *y* just in case *x* is an agent with some power or authority and . . . *y* is suffering unjustly or wrongfully under *x* or as a result of *x's* unjust exercise of power.
>
> (Haslanger 2004, 98–9)

Examples of agent oppression are familiar enough: these include both a tyrant who exercises his power over his subjects, and a rapist who does the same to his target. As Haslanger goes on to note, however, many feminist theorists have been concerned primarily not with agent oppression, but with what we might call 'structural oppression'.

Marilyn Frye describes an oppressive *structure* as "part of an enclosing structure of forces and barriers which tends to the immobilization and reduction of a group or category of people" (Frye 1983, 10–11). These structures—the "forces and barriers"—can include institutions, laws, public policies, and norms, among other things. They create restrictions, and thereby disadvantage those who are subject to them.

But not every disadvantage is oppressive. The structures that create poor, racially segregated neighborhoods in Detroit might disadvantage both white people (by making it uncomfortable and unsafe for them to visit those neighborhoods) and people of color (by making it difficult for them to live elsewhere). But both groups are not thereby oppressed.

What more then is required for (structural) oppression? Several theorists have argued that it requires not mere disadvantage, but also *unjust* and *systematic* disadvantage. Frye emphasizes the latter aspect in the following passage:

> The experience of oppressed people is that the living of one's life is confined and shaped by forces and barriers which are not accidental or occasional and hence avoidable, but are systematically related to each other in such a way as to catch one between and among them and restrict or penalize motion in any direction. It is the experience of being caged in: all avenues, in every direction, are blocked or booby trapped.
>
> (Frye 1983, 4)

To understand the full significance of such oppression, Frye argues, we need to take a 'macroscopic' view, to see how various oppressive structures interact to affect every aspect of the lives of oppressed groups.

The notion of structural oppression has been further elaborated by Iris Marion Young, who argues that it can take several forms (Young 1990, chapter 2). She distinguishes five such forms, or 'faces', of oppression: exploitation, marginalization, powerlessness, cultural imperialism, and violence. Young takes none of these forms to be more explanatorily basic than the others. Other theorists have mentioned further forms of oppression: Sandra Bartky, for example, discusses *psychological* oppression, while Fricker considers *epistemic* oppression (Bartky 1990, chapter 2, Fricker 1999, 2007). We can ask how these relate to the five faces Young discusses, and whether there are other distinct faces beyond these.

Young, like Frye, argues that structural oppression doesn't require an oppressor, at least in the sense that it needn't be the result of any agent's intent to oppress (or even disadvantage) the groups in question. Rather, such oppression can be the unintended consequence of actions undertaken with quite benign intentions. To capture this, we might distinguish between those who benefit from a particular form of oppression—call them 'the privileged group'—and oppressors. Structural oppression creates privileged groups, but doesn't require an oppressor. In this respect, it is arguably different from agent oppression.

Developing these ideas, several philosophers have emphasized that actions that look relatively innocuous taken alone might nevertheless together ramify into an oppressive structure (Haslanger 2004, McGowan 2009, among others). This point is double-edged: on the one hand, it means that oppression may be harder to identify and more pervasive than previously understood; on the other hand, it also means that there are more possibilities for disrupting oppression than previously recognized.

Drawing on several of the antecedents mentioned above, Ann Cudd has offered what she describes as a "social force analysis of oppression" (Cudd 2006, 22). Cudd suggests several desiderata for an adequate theory of oppression, including the following two: such a theory must (i) explain how oppression endures over time, and (ii) point towards a solution, that is, towards some means for mitigating that oppression.

For Cudd, oppression exists in any context in which the following four conditions are satisfied:

1) *The harm condition*: There is a harm that comes out of an institutional practice.
2) *The social group condition*: The harm is perpetrated through a social institution or practice on a social group whose identity exists apart from the oppressive harm in 1).
3) *The privilege condition*: There is another social group that benefits from the institutional practice in 1).

> 4) *The coercion condition*: There is unjustified coercion or force that brings about the harm.
>
> (Cudd 2006, 25)

Thus, for Cudd, oppression comes primarily out of institutions. It is maintained by social forces of different kinds: forces of violence, as well as economic and psychological forces. These forces work together to make it *rational*, at least in the short term, for the oppressed to choose actions that, in the long term, end up perpetuating the oppression of their own group. That, Cudd argues, plays an important role in explaining how oppression endures, and so must be taken into account in thinking about how to mitigate it. (See Allen 2008 for criticisms of Cudd's view.)

Thus far, I've focused on attempts to theorize oppression as such. But as we saw earlier, hooks's characterization of feminism makes mention of a particular form of oppression, namely, *sexist* oppression. Sexist oppression, or oppression on the basis of gender, has to be distinguished from oppressions based on race, class, disability, and other factors. But it won't do to say that oppression is sexist just in case it results in disadvantage to women, because all of the forms just mentioned share that characteristic.

We can make a start on this problem by saying, with Frye, that each of the forms of oppression mentioned above targets certain individuals for "application of oppressive pressures" in virtue of their membership in a particular social group, whether that be the group of women, or blacks, or the disabled, or some other (Frye 1983, 15). Then, sexist oppression is that oppression which targets individuals in virtue of their membership in the group 'women'; or, to put it another way, it's the oppression which targets certain individuals *as women*.

Of course, this is only the beginnings of a solution, since it remains to be specified how the 'in virtue of membership in group *G*' and 'as *X*' locutions are to be understood here. And this proves to be a difficult problem, for several reasons. One major reason is related to the worry about exclusion mentioned in §2: if what it is to *be* a woman differs across historical eras, races, sexualities, classes, disabilities, and nationalities, we should expect what it is to be *oppressed* as a woman to similarly vary across these dimensions (Spelman 1988). That raises the worry that any substantive account of targeting individuals for oppression as women will end up excluding some of what should count as sexist oppression. (See Haslanger 2004 for a response.)

In addition, even given an adequate account of what distinguishes the various forms of oppression, there is the further crucial question about how these forms interact. That question is of paramount importance for understanding the oppression faced by women of color, for example. Some theorists have argued that understanding this requires theorizing how multiple oppressions interact to transform, and exacerbate, the effects of each (Crenshaw 1991). Considering each form of oppression on its own will fail to do justice to the experiences of those at their intersections. (See Lugones 2003 and Alcoff 2006 for more on this problem of intersectionality; see also Garry 2011 and Sheth 2014 for some notable recent work.)

Much of the feminist work on oppression discussed in this section emphasizes its structural aspects. Though agent oppression may count as a genuine form of oppression, many feminist theorists have thought that focusing merely on agent oppression misses out on the deepest problems facing women. A second recurring theme has been the importance of

theorizing not merely oppression, but also, possibilities for resistance. Insofar as feminism aims not just to uncover sexist oppression, but also to remedy it, making room for these possibilities is a crucial part of the project.

4.1 Objectification

As mentioned in §1, the notion of objectification is one that feminist theorists have borrowed from others, notably Kant, and then developed in distinctive ways.

Centrally, objectification involves treating someone (or something) *as* an object. Treating another as an instrument for one's own purposes is one way of objectifying them. For Kant, that constitutes an infringement of the moral requirement that humanity be treated as ends, never as means only (Kant 1785/2012). Kant worries about sexual activity outside the context of monogamous marriage, since, on his view, such activity involves treating others as mere instruments for our own sexual gratification (Kant 1762–1794/1997).

The Kantian notion of objectification finds echoes in the work of feminist theorists like Andrea Dworkin and Catharine MacKinnon (Dworkin 1981, MacKinnon 1987; see also Herman 1993 for discussion of the Kantian echoes). Both MacKinnon and Dworkin argue that the *sexual* objectification of women is central to women's oppression. For MacKinnon, in societies like ours, women as women are constructed as sex objects, while men as men are constructed as objectifiers. (The 'as women' and 'as men' qualifiers are important here, for it is compatible with this view that *some* women objectify, and *some* men are objectified.) As sex objects, women are available for use for sexual gratification.

Pornography, on MacKinnon's view, plays a central role in this construction of women's (and men's) sexuality:

> Pornography *participates* in its audience's eroticism through creating an accessible sexual object, the possession and consumption of which *is* male sexuality, as socially constructed; to be consumed and possessed as which, *is* female sexuality, as socially constructed . . . Pornography defines women by how we look according to how we can be sexually used.
>
> (MacKinnon 1987, 173, original emphasis)

These views about pornography, sexual objectification, and women's oppression have been enormously influential within feminist work, and also enormously controversial. (See Haslanger 1993, Hornsby 1993, Langton 1993, and McGowan 2005 for some interpretations of MacKinnon's and Dworkin's claims. See Butler 1997, Green 2000, and Shrage 2005 for criticisms.)

Drawing upon both the Kantian and feminist antecedents, Martha Nussbaum has offered an analysis of the concept 'objectification' that tries to make sense of the various claims (Nussbaum 1995). She distinguishes seven distinct ways in which something (or someone) can be treated as an object:

1) *Instrumentality:* The objectifier treats the object as a tool of his or her purposes;
2) *Denial of autonomy:* The objectifier treats the object as lacking in autonomy and self-determination;

3) *Inertness:* The objectifier treats the object as lacking in agency, and perhaps also in activity;
4) *Fungibility:* The objectifier treats the object as interchangeable (a) with other objects of the same type, and/or (b) with objects of other types;
5) *Violability:* The objectifier treats the object as lacking in boundary-integrity, as something that it is permissible to break up, smash, break into;
6) *Ownership:* The objectifier treats the object as something that is owned by another, can be bought or sold, etc.;
7) *Denial of subjectivity:* The objectifier treats the object as something whose experiences and feelings (if any) need not be taken into account.

(Nussbaum 1995, 257)

Because 'objectification' is a cluster concept on this view, none of these features is *necessary* for it to apply, though sufficiently many of them will generally suffice. Nussbaum also emphasizes that objectification, including sexual objectification, need not be morally problematic; rather, it is sometimes a healthy part of sexual life. The problem, then, is to distinguish between objectification that is morally problematic and objectification that is not, in a principled way.

Rae Langton takes Nussbaum's list as a starting point, but thinks that it fails to capture some features central to the feminist understanding of objectification (Langton 2009, chapter 10). She adds three entries to Nussbaum's list:

8) *Reduction to body*: [The objectifier treats the object] as identified with its body, or body parts.
9) *Reduction to appearance:* [The objectifier treats the object] primarily in terms of how it looks, or how it appears to the senses.
10) *Silencing*: [The objectifier treats the object] as silent, lacking the capacity to speak.

(Langton 2009, 228–9)

Langton augments Nussbaum's view in some further ways. First, she emphasizes that theorizing objectification requires paying attention both to the 'object' aspect, that is, what it is to be an object, or object-like, as well as to the 'treating as' aspect (Langton 2009, chapter 10). Nussbaum's list, with the additions above, focuses on the first aspect. Langton draws on a Humean notion of projection to illuminate the second aspect, distinguishing several projective mechanisms that can help objectification happen, as well as hide that it is happening (Langton 2004).

Second, Langton also distinguishes the moral and the epistemic dimensions of objectification: on her view, an objectifier makes *both* a moral and an epistemic error. The moral dimension is familiar from the Kantian prohibition on treating others as means only, mentioned at the beginning of this section. The epistemic dimension is less familiar, but, for Langton, no less important: it involves failing to notice the objectifier's role in bringing about what is falsely taken to be natural, or at least, independent of the objectifier's actions. Langton finds both dimensions in MacKinnon's work on sexual objectification:

> Sexual objectification is partly a matter of the 'possession' and 'use' of women, [MacKinnon] says; and it is at the same time 'an elaborate projective system'. . . . [A] propensity to project properties onto women might help make those properties real, in circumstances where, as MacKinnon puts it, 'the world actually arranges itself to affirm what the powerful want

> to see'. Projection of sexual submissiveness, for example, might help make women sexually submissive.
>
> (Langton 2009, 12, quoting MacKinnon 1989, 140–1 and MacKinnon 1987, 164)

(See Haslanger 1993 and McGowan 2005 for more on MacKinnon and the epistemic error involved in objectification.)

Theorists have also paid attention to particular *kinds* of objectification. Sexual objectification has already figured prominently in this discussion. Langton identifies silencing of speakers as another kind of objectification; she and others have articulated what constitutes silencing in the relevant sense (Langton 1993, Hornsby 1993, West 2003). Fricker has discussed what she calls 'epistemic objectification' in the context of characterizing the wrong done to testifiers who are unjustly excluded from the realm of epistemic agency (Fricker 2007, chapter 6).

Feminist theories of objectification thus cover a lot of ground, focusing especially, but not only, on sexual objectification. A recurring concern in these discussions has to do with revealing that objectification is involved at all, that is, that what looks at first glance like an aspect of women's (or men's) natures is actually the result of exercises of social power. As such, what was assumed to be natural is shown to be deeply social.

5. Situated Knowledge (and Situated Agency)

"The central concept of feminist epistemology," writes Elizabeth Anderson, "is that of a situated knower, and hence of situated knowledge" (Anderson 2012, 2). A situated knower here is someone who occupies a specific situation or perspective, where that situation or perspective affects what she knows (or doesn't know), or *how* she knows it. Situated knowledge, then, is just knowledge had by a situated knower.

Examples of situated knowledge are familiar enough: I have first-personal knowledge of some things (for example, that I have a caffeine headache coming on soon) and third-personal knowledge of others. Because I am sitting at this table at this café, I can see the seam running down the middle of its surface; the person sitting on the other side of the café cannot see that detail, or at least, cannot see it in the same way. Because I want to finish writing this chapter, I experience the noise of lorries passing by in the alley outside as irritating. As Anderson points out, all of these—first- vs. third-personal knowledge, embodied knowledge, interest- or emotion-laden knowledge—are examples of situated knowledge (Anderson 2012).

Feminist epistemologists have been particularly concerned with one kind of social situation, namely, gender, in all its aspects. (Recall that gender has many aspects, including gender norms, gender symbolism, gender identity, and so on.) Thus, such theorists have been concerned with the way in which a knower's gender, in any of its aspects, affects what she knows, and *how* she knows it. Let's call this kind of situated knowledge 'gendered knowledge'. Much of the work discussed in §1 of this chapter focuses precisely on how knowledge is gendered in this sense.

As emphasized in §1, concern with gendered knowledge has both a descriptive and a normative aspect. The descriptive aspect tries to capture what role gender in fact plays in what, and how, a knower knows. The normative aspect focuses on what role gender should play in what, and how, a knower knows. As I also mentioned in §1, many feminist epistemologists have argued that gender is regularly a source of unjust epistemic disadvantage, and as such, its impact should be reduced; but some have also argued that gender is sometimes a source of epistemic privilege. Insofar as that privilege is, on some theories, tied to women's situation as targets of oppression, it is not obviously a cause for celebration.

On Anderson's rendering, feminist epistemology with its central focus on an aspect of knowers' social situations turns out to be a branch of *social* epistemology. (But see Rooney 2012 and Grasswick 2013 for some doubts about whether that rendering captures all there is to feminist epistemology.)

Because I have discussed feminist epistemology in some detail already, I won't say much more here. Instead, I will finish this section by arguing that this concern with situatedness—especially, but not only, its gender aspects—is also a recurring theme in other areas of feminist work in philosophy.

We can model a notion of a situated (moral) agent on that of the situated knower described above. On this conception, a situated (moral) agent would be one who occupies a particular situation or perspective, where that situation or perspective affects her morally significant attributes. Those attributes include what she is morally permitted to do, what she is morally required to do, what she has moral reason to do, how she is harmed, how she is morally wronged, and so on.

Some central debates in feminist ethics are concerned with moral agents' social situations, especially their gender. Consider, for instance, debates about autonomy. Several feminist philosophers have rejected older notions of autonomy precisely because they fail to take into account agents' social situations, including their gender (Mackenzie and Stoljar 2000, Introduction). One line of criticism has been that these older notions tend to unfairly discount as non-autonomous women who are deeply embedded in caring relationships, either as caregiver or cared-for. A second line of criticism has held that these notions fail to notice how sexist (and other forms of) oppression can work to covertly undermine their targets' autonomy. In response to these worries, theorists have tried to develop *relational* notions of autonomy, which aim to

> analyze the implications of the intersubjective and social dimensions of selfhood and identity for conceptions of individual autonomy and moral and political agency.
>
> (Mackenzie and Stoljar 2000, 4)

Thus, we might say that these relational notions try to capture how the situatedness of moral agents, including their gender, matters for their status as autonomous agents. (See Oshana 2006 and Westlund 2009 for some recent relational theories.)

6. Themes

I've now discussed several tools that have been central to feminist work in philosophy, and tried to indicate why they have been so important. In closing, I'll briefly mention some themes that have come up repeatedly in the discussion in this chapter:

1) Focus on the social: Much feminist work has been aimed at showing that what had been considered natural, or biological, is in in fact social, though no less real on that count. This kind of work doesn't necessarily deny the significance of biology, or of biological constraints. Instead, it urges caution about what we take to be natural, or biological, and on what grounds.
2) Focus on the structural: Feminist philosophers have been concerned with the roles that social structures play in our lives, and particularly with the ways in which those structures interact to limit our options.
3) Draw upon work from other disciplines: Because feminist philosophers are interested in the role of gender in various practices, work in a number of fields outside philosophy is highly relevant to our concerns; this includes biology, history, legal theory, science studies, and sociology, among other fields. As such, feminist work in philosophy is not autonomous vis-à-vis these other disciplines. Rather, it draws regularly on work in those disciplines (and, in turn, can influence that work).
4) Aim for inclusion: Insofar as feminism is concerned with gender as such, or women *as* women, it is crucial that generalizations include all women, not merely women of the theorizer's acquaintance. Given the demographics of professional philosophy, that means that armchair philosophizing—or, for that matter, faculty lounge philosophizing—is unlikely to suffice for this purpose. (See Gines 2011 and Paxton, Figdor, and Tiberius 2012 for more on those demographics.)
5) Begin with the diversity of actual world cases: Feminist theorizing in philosophy is often driven by questions arising from actual cases. Though thought experiments have a role to play as well, work in this area is generally aimed at capturing the diversity of those actual cases.
6) Focus on situatedness: Much feminist philosophy focuses on women's particular situations, and on how those situations affect them as knowers, and as agents more generally.
7) Engage in a political enterprise: Insofar as it aims to uncover and mitigate sexism, feminist work in philosophy is not merely a theoretical, but also a political enterprise. As such, it is a deeply value-laden inquiry.
8) Make room for agency and resistance: As just mentioned, feminist work in philosophy aims not merely to uncover sexism, but also to suggest remedies. Some theorists have suggested that feminist theorizing that fails to point the way towards resisting and mitigating sexism is pointless.

The themes just mentioned are not the only ones that recur through feminist philosophy. Others—for example, taking a macroscopic view as urged by Frye, or making room for complexity of relationships as urged by Helen Longino—arguably belong on this list as well

(Frye 1983, Longino 1995). Nonetheless, I take this list to be a significant, even if partial, catalog of methodological inclinations within feminist work in philosophy.[7]

References

Alcoff, Linda Martín. 2006. *Visible Identities: Race, Gender, and the Self*. Oxford: Oxford University Press.

Alcoff, Linda Martín. 2010. Epistemic identities. *Episteme* 7(2): 128–37.

Allen, Amy. 2008. Rationalizing oppression. *Journal of Power* 1(1): 51–65.

Anderson, Elizabeth. 2004. Uses of value judgments in science: A general argument, with lessons from a case study of feminist research on divorce. *Hypatia* 19(1): 1–24.

Anderson, Elizabeth. 2012. Feminist epistemology and philosophy of science. In *The Stanford Encyclopedia of Philosophy*. URL=<http://plato.stanford.edu/archives/fall2012/entries/feminism-epistemology/>. Accessed September 28, 2015.

Antony, Louise. 2012. Is there a "feminist" philosophy of language? In Crasnow and Superson (2012), 245–85.

Antony, Louise M. and Charlotte E. Witt (eds.). 1993/2002. *A Mind of One's Own: Feminist Essays on Reason and Objectivity*, 1st edition/2nd edition. Boulder, CO: Westview Press.

Barnes, Elizabeth. 2014. Going beyond the fundamental: Feminism in contemporary metaphysics. *Proceedings of the Aristotelian Society* 114(3pt3): 335–51.

Bartky, Sandra L. 1990. *Femininity and Domination: Studies in the Phenomenology of Oppression*. New York: Routledge.

Butler, Judith. 1997. *Excitable Speech: A Politics of the Performative*. New York: Routledge.

Code, Lorraine. 1987. *Epistemic Responsibility*. Hanover, NH: University Press of New England.

Code, Lorraine. 1991. *What Can She Know?: Feminist Theory and the Construction of Knowledge*. Ithaca, NY: Cornell University Press.

Collins, Patricia Hill. 1990. *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. Boston: Unwin Hyman.

Crasnow, Sharon L. and Anita M. Superson (eds.). 2012. *Out from the Shadows: Analytical Feminist Contributions to Traditional Philosophy*. Oxford: Oxford University Press.

Crenshaw, Kimberlé. 1991. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review* 43(6): 1241–99.

Cudd, Ann E. 2006. *Analyzing Oppression*. Oxford: Oxford University Press.

Cudd, Ann E. 2012. Resistance is (not) futile: Analytical feminism's relation to political philosophy. In Crasnow and Superson (2012), 15–31.

Dotson, Kristie. 2012. A cautionary tale: On limiting epistemic oppression. *Frontiers* 33(1): 24–47.

Dworkin, Andrea. 1981. *Pornography: Men Possessing Women*. New York: Perigee Books.

Fausto-Sterling, Anne. 1993. The five sexes: Why male and female are not enough. *The Sciences* 33(2): 20–4.

Fausto-Sterling, Anne. 2000. *Sexing the Body: Gender Politics and the Construction of Sexuality*. New York: Basic Books.

[7] I am very grateful to Mary Kate McGowan and Brian Weatherson for helpful comments on earlier drafts.

Ferguson, Ann. 1983. On conceiving motherhood and sexuality: A feminist materialist approach. In Joyce Trebilcot (ed.), *Mothering: Essays in Feminist Theory*, 153–82. Totowa, NJ: Rowman and Allanheld.

Fricker, Miranda. 1999. Epistemic oppression and epistemic privilege. *Canadian Journal of Philosophy* 29(sup1): 191–210.

Fricker, Miranda. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press.

Fricker, Miranda and Jennifer Hornsby (eds.). 2000. *The Cambridge Companion to Feminism in Philosophy*. Cambridge: Cambridge University Press.

Frye, Marilyn. 1983. *The Politics of Reality: Essays in Feminist Theory*. Trumansburg, NY: Crossing Press.

Garry, Ann. 2011. Intersectionality, metaphors, and the multiplicity of gender. *Hypatia* 26(4): 826–50.

Gines, Kathryn T. 2011. Being a black woman philosopher: Reflections on founding the Collegium of Black Women Philosophers. *Hypatia* 26(2): 429–37.

Glasgow, Joshua. 2009. *A Theory of Race*. New York: Routledge.

Grasswick, Heidi. 2013. Feminist social epistemology. In *The Stanford Encyclopedia of Philosophy*. URL=<http://plato.stanford.edu/archives/spr2013/entries/feminist-social-epistemology/>. Accessed October 19, 2015.

Green, Leslie. 2000. Pornographies. *Journal of Political Philosophy* 8(1): 27–52.

Harding, Sandra. 1991. *Whose Science? Whose Knowledge?: Thinking from Women's Lives*. Ithaca, NY: Cornell University Press.

Harding, Sandra and Merrill B. Hintikka (eds.). 1983. *Discovering Reality: Feminist Perspectives on Epistemology, Metaphysics, Methodology, and Philosophy of Science*. Dordrecht: D. Reidel.

Hartsock, Nancy C. M. 1987. The feminist standpoint: Developing the ground for a specifically feminist historical materialism. In Sandra Harding (ed.), *Feminism and Methodology: Social Science Issues*, 157–80. Bloomington, IN: Indiana University Press.

Haslanger, Sally. 1993. On being objective and being objectified. In Antony and Witt (1993), 85–125. Reprinted in her (2012), 35–82.

Haslanger, Sally. 2000. Gender and race: (What) are they? (What) do we want them to be? *Noûs* 34(1): 31–55. Reprinted in her (2012), 221–47.

Haslanger, Sally. 2003-4. Future genders? Future races? *Philosophic Exchange* 34(1): 5–27. Reprinted in her (2012), 248–72.

Haslanger, Sally. 2004. Oppressions: Racial and other. In Michael P. Levine and Tamas Pataki, *Racism in Mind*, 97–123. Ithaca, NY: Cornell University Press. Reprinted in her (2012), 311–38.

Haslanger, Sally. 2012. *Resisting Reality: Social Construction and Social Critique*. Oxford: Oxford University Press.

Herman, Barbara. 1993. Could it be worth thinking about Kant on sex and marriage? In Antony and Witt (1993), 49–67.

hooks, bell. 2000. *Feminist Theory: From Margin to Center*, 2nd edition. Cambridge, MA: South End Press.

Hornsby, Jennifer. 1993. Speech acts and pornography. *Women's Philosophy Review* 10: 38–45.

Jones, Karen. 2002. The politics of credibility. In Antony and Witt (2002), 154–76.

Kant, Immanuel. 1762–1794/1997. *Lectures on Ethics*. Edited by Peter Heath and J. B. Schneewind. Translated by Peter Heath. Cambridge: Cambridge University Press.

Kant, Immanuel. 1785/2012. *Groundwork of the Metaphysics of Morals*, revised edition. Translated and edited by Mary Gregor and Jens Timmermann. Cambridge: Cambridge University Press.

Langton, Rae. 1993. Speech acts and unspeakable acts. *Philosophy & Public Affairs* 22(4): 293–330. Reprinted in her (2009), 25–63.

Langton, Rae. 2004. Projection and objectification. In Brian Leiter (ed.), *The Future for Philosophy*, 285–303. Oxford: Oxford University Press. Reprinted in her (2009), 241–66.

Langton, Rae. 2009. *Sexual Solipsism: Philosophical Essays on Pornography and Objectification*. Oxford: Oxford University Press.

Longino, Helen. 1995. Gender, politics, and the theoretical virtues. *Synthese* 104(3): 383–97.

Lugones, María. 2003. *Pilgrimages/Peregrinajes: Theorizing Coalition Against Multiple Oppressions*. Lanham, MD: Rowman & Littlefield.

McGowan, Mary Kate. 2005. On pornography: MacKinnon, speech acts, and "false" construction. *Hypatia* 20(3): 22–49.

McGowan, Mary Kate. 2009. Oppressive speech. *Australasian Journal of Philosophy* 87(3): 389–407.

Mackenzie, Catriona and Natalie Stoljar (eds.). 2000. *Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Self*. Oxford: Oxford University Press.

MacKinnon, Catharine A. 1987. *Feminism Unmodified: Discourses on Life and Law*. Cambridge, MA: Harvard University Press.

MacKinnon, Catharine A. 1989. *Toward a Feminist Theory of the State*. Cambridge, MA: Harvard University Press.

Medina, José. 2013. *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and Resistant Imaginations*. Oxford: Oxford University Press.

Mikkola, Mari. 2011. Ontological commitments, sex and gender. In Witt (2011a), 67–83.

Mikkola, Mari. 2012. Feminist perspectives on sex and gender. In *The Stanford Encyclopedia of Philosophy*. URL=<http://plato.stanford.edu/archives/fall2012/entries/feminism-gender/>. Accessed October 19, 2015.

Moi, Toril. 1999. *What is a Woman?: And Other Essays*. Oxford: Oxford University Press.

Nussbaum, Martha. 1995. Objectification. *Philosophy & Public Affairs* 24(4): 249–91.

Ortega, Mariana. 2006. Being lovingly, knowingly ignorant: White feminism and women of color. *Hypatia* 21(3): 56–74.

Oshana, Marina. 2006. *Personal Autonomy in Society*. Aldershot, UK: Ashgate.

Paxton, Molly, Carrie Figdor, and Valerie Tiberius. 2012. Quantifying the gender gap: An empirical study of the underrepresentation of women in philosophy. *Hypatia* 27(4): 949–57.

Richardson, Sarah S. 2010. Feminist philosophy of science: History, contributions, and challenges. *Synthese* 177(3): 337–62.

Rooney, Phyllis. 2012. What is distinctive about feminist epistemology at 25? In Crasnow and Superson (2012), 339–75.

Saul, Jennifer. 2006. Philosophical analysis and social kinds: Gender and race. *Proceedings of the Aristotelian Society*, Supplementary Volume 80(1): 119–43.

Saul, Jennifer. 2012. Politically significant terms and philosophy of language: Methodological issues. In Crasnow and Superson (2012), 195–216.

Schiebinger, Londa. 2004. Feminist history of colonial science. *Hypatia* 19(1): 233–54.

Sheth, Falguni A. 2014. Interstitiality: Making space for migration, diaspora, and racial complexity. *Hypatia* 29(1): 75–93.

Shrage, Laurie. 2005. Exposing the fallacies of anti-porn feminism. *Feminist Theory* 6(1): 45–65.
Spelman, Elizabeth V. 1988. *Inessential Woman: Problems of Exclusion in Feminist Thought*. Boston: Beacon Press.
Sveinsdóttir, Ásta K. 2012. Review of *The Metaphysics of Gender. Notre Dame Philosophical Reviews*. URL=<https://ndpr.nd.edu/news/30682-the-metaphysics-of-gender/>. Accessed September 28, 2015.
Thomson, Judith Jarvis. 1971. A defense of abortion. *Philosophy & Public Affairs* 1(1): 47–66.
Townley, Cynthia. 2006. Toward a revaluation of ignorance. *Hypatia* 21(3): 37–55.
Townley, Cynthia. 2011. *A Defense of Ignorance: Its Value for Knowers and Roles in Feminist and Social Epistemologies*. Lanham, MD: Lexington Books.
Tuana, Nancy. 2006. The speculum of ignorance: The women's health movement and epistemologies of ignorance. *Hypatia* 21(3): 1–19.
West, Caroline. 2003. The free speech argument against pornography. *Canadian Journal of Philosophy* 33(3): 391–422.
Westlund, Andrea C. 2009. Rethinking relational autonomy. *Hypatia* 24(4): 26–49.
Witt, Charlotte (ed.). 2011a. *Feminist Metaphysics: Explorations in the Ontology of Sex, Gender and the Self*. Dordrecht: Springer.
Witt, Charlotte. 2011b. *The Metaphysics of Gender*. Oxford: Oxford University Press.
Wylie, Alison. 2003. Why standpoint matters. In Robert Figueroa and Sandra Harding (eds.), *Science and Other Cultures: Issues in Philosophies of Science and Technology*, 26–48. New York: Routledge.
Young, Iris Marion. 1990. *Justice and the Politics of Difference*. Princeton, NJ: Princeton University Press.

CHAPTER 36

CRITICAL PHILOSOPHY OF RACE

CHARLES MILLS

1. Introduction

Critical philosophy of race is distinguished from traditional—*uncritical*—philosophy of race in being multiply "critical." It is critical of racism, as ideas, beliefs, and values, as well as social institutions and practices; it is generally, at least in recent decades, critical also of the traditional naturalistic understanding of race; and it is critical of the denial of the past and ongoing significance of race to the making of modernity and the contemporary world. The analogy with feminist theory, while not perfect, is a useful one. Traditional gender theory has historically been sexist and has naturalized gender and gender relations. Feminism can then be seen as "critical" gender theory in the sense that it denaturalizes gender relations, viewing them as social artifacts that usually subordinate women. Similarly, most race theory over the time period race has existed as a concept has been racist, and has naturalized race and race relations. Critical race theory is then like feminism in challenging historically dominant race theory and its conceptions of race, and in seeking a society of racial equality and perhaps, one day (though this is more controversial), one without races at all. The term was originally coined in the 1980s by American legal theorists, and meant specifically to refer to critical race theory in the law (Crenshaw et al. 1995). But as with most other intellectual movements, the phenomenon itself long predates its christening, so that one can speak of critical race theory and critical philosophy of race *avant la lettre*. In the black diaspora, which has played a crucial role in developing critical race theory and critical race philosophy, credit is usually given to such pioneering figures as Anténor Firmin (2002) (Haiti), W. E. B. Du Bois (Sundquist 1996) (United States), and Frantz Fanon (2004, 2008) (Martinique), while José Vasconcelos (1997) (Mexico) is a major contributor in the Latin world.

Critical philosophy of race is, then, the term recently proposed to designate critical race theory in the discipline of philosophy specifically.[1] What demarcates critical philosophy of

[1] See the new journal from Penn State University Press, *Critical Philosophy of Race*, launched in

race is thus its subject matter and the approach to it, not the racial identity of its practitioners. Native American, Asian, African, and Australian aboriginal philosophy will predominantly be the work of people of color, but that does not make them examples of critical philosophy of race. It is the critical thematization of race that is defining, and at least by the standard narrative (though it is contested, as we will soon see), race would not have been a relevant concept for these populations prior to European contact. "Race" in its original conception here denotes natural sub-sections of the human race, morphologically distinct, and linked by ancestry to particular geographical regions, which are taken to have differentiated biological, psychological, and cultural traits which are innate and invariant, as against ethnic groups into which outsiders can culturally assimilate.

However, as the field is still in the process of being mapped and formalized, many other boundaries remain fuzzy. For example, does one have to see race as central to the making of the modern world to be a critical race philosopher, or is this not a prerequisite? What about eighteenth- and nineteenth-century theorists who opposed racism and argued for racial equality, but did believe that races were indeed natural biological sub-sections of the human race, and not, in a contemporary vocabulary, social constructs? Or thinkers of color who, in the reaction against white domination, sometimes lapsed into "oppositional" racial chauvinisms of their own? Should they count as critical race philosophers, or are they insufficiently "critical"? These questions and many others remain to be worked out.

Finally, in trying to establish a comprehensive genealogy of critical philosophy of race, it will be necessary to seek out a wide variety of texts in addition to the scholarly essay or monograph, such as fiction and poetry, material that is oral as well as written, and produced by people who generally have no formal philosophical training, but whose work is nonetheless "philosophical" in the sense that it explores, from the perspective of race, distinctively philosophical issues. Moreover, because of the novelty of this subject as a formalized philosophical perspective, much will need to be done to discover and translate examples of critical philosophy of race in oppositional anti-colonial, anti-imperialist, and anti-racist work from the non-Anglo world. *Caveat lector*, then: the following account is biased by the Anglo-American archive and those texts translated into it.[2]

2. Periodization

Though they have many commonalities, a crucial disanalogy between race and gender as social categories and social realities is that gender relations go back to the origin of the species, whereas race is, at least arguably, more recent. All human communities have, in one form or another, constructed gender, but only comparatively recently has race unequivocally become a marker of human social division. Scholarly controversy exists on how to think of race, the date of its emergence as a recognized social category, its social origin,

2013, and Paul Taylor's (2012) useful introduction to his edited, four-volume collection of reprints of some classic pieces in the field.

[2] Robert Bernasconi (2012) should be consulted for his valuable overview from the Continental (phenomenological) perspective, which has received inadequate attention here. See also Bernasconi and Cook (2003) and Gines (2012).

and the causes of its genesis—the what, when, where, and why of race (Boxill 2001a). For traditional, *uncritical* race theory, race was natural, so answering these questions was of course much easier. But we now view these answers as wrong. See, for example, Arthur de Gobineau's *The Inequality of Human Races* (1999), for which racial types are "absolutely fixed, hereditary, and *permanent*" (125), and divided by "irreconcilable antagonisms" (179) that have the character of "natural laws" (133). Humanity is a "vast hierarchy" of "unlike and unequal parts" (181), consisting of (reading from bottom to top) "the black, the yellow, and the white" (205), with "all civilizations deriv[ing] from the white race" (210). (Anténor Firmin [2002] produced the most impressively detailed and thorough nineteenth-century (1885) refutation of Gobineau, but unfortunately his own book never achieved the racist text's wide translation and global circulation.)

George Fredrickson (2002) articulates what has been the most influential position in recent decades: race—racial categories, race-thinking—is a product of the modern world or, at the earliest, the late medieval period, and is largely a creation of the West. If a single place and time could be chosen as neatly epitomizing this transition from a non-racial to a racial world, it is the Spain of 1492: the year of the completion of the *Reconquista* from the Moors, the expulsion of the Jews, and the Columbian embarkation on the missions of conquest that would eventually lead to a globally dominant European racial system. The invocation of the category of *limpieza de sangre* to disqualify as genuine believers those Jews who became *conversos* to escape the persecution of the Inquisition signaled the shift from a religious anti-Judaism to a racial anti-Semitism, while the "discovery" of a continent not described in the Bible gave rise to the challenge of developing a taxonomy to locate these "new" humans. As Lewis Gordon (2008) has argued, the issue of philosophical anthropology, of how best to conceptualize and understand the human and differentiate it from the non-human, becomes especially crucial in this period of the European mapping of the world and its inhabitants. Race increasingly functions as a line of essential demarcation that separates the unambiguous humans ("white") from the question-mark humans ("non-white"), those human in appearance (more or less) but inferior in behavior and/or culture and/or biology. On this analysis, then, earlier prejudicial depictions of Jews and (what we now see as) people of color in early medieval and classical Western texts should not be categorized as racism but rather as ethnocentrism, religious bigotry, xenophobia, or color prejudice (Hannaford 1996). The periodization of race and racism really begins with modernity.

However, a dissenting narrative has long existed that would push the timeline back much further (Gossett 1997; Delacampagne 1983). These scholars argue that while medieval and ancient Western conceptions of race were admittedly different from modern ones (the what), they do nonetheless count as conceptions of *race*, and not just ethnicity, religion, nationality, or color. In addition, some theorists have contended that it is a mistake to focus exclusively on the West. While it is undeniable that racism achieves its most influential ("world-historical") and poisonous flowering there, other medieval and ancient civilizations, it is claimed, have also revealed nascent or well-developed patterns of early racial thought: China, India, Japan, the Arab world (Hund 2006).

The case for race as pre-modern has been greatly strengthened in the past decade by Benjamin Isaac's massive *The Invention of Racism in Classical Antiquity* (2004), which explicitly sets out to challenge the conventional narrative, and by a follow-up conference

volume edited by Isaac and others, *The Origins of Racism in the West* (Eliav-Feldon, Isaac, Ziegler 2009). These two books argue that racism does indeed go back to antiquity—Isaac spoke of "proto-racism" in the first book, but drops the qualifying prefix as potentially misleading by the time of the second one—but that, contra claims of multiple origins, it is emphatically Western. His reasons have, ironically enough, led to the charge of racism being leveled against *him*. Only in ancient Greece, he and his co-editors claim, not in any of the other civilizations of the classical world, do we find the degree of theoretical sophistication necessary for unthinking bigotry to rise to the level of abstraction that would justify the designation *racism*:

> [T]he opinion that racism needs to be discussed also in respect of non-western cultures rests on a confusion of racism and other forms of prejudice. The mythology and sagas of numerous peoples contain stories about their own separateness and superiority. However, prejudice expressed as myth is radically different from prejudice elevated to the level of scientific truth. . . . We do not assume that prejudice and bigotry were invented in the West; we claim rather that the specific form of rationalizing these prejudices and attempting to base them on systematic abstract thought was developed in antiquity and taken over in early modern Europe. . . . It is generally accepted that Greek civilization was the first to raise abstract, systematic thought to a level that we now recognize as approaching our own. They were the first to develop abstract concepts in their thinking about nature and to systematize those ideas. (2009, 9)

Adjudicating this controversy cannot be done here. But it does raise the possibility of the theoretical object of critical philosophy of race being much older than standardly thought, and (for those who disagree with Isaac et al.) arising in contestation with non-Western forms of racism as well as the more familiar Western variety. However, the key point we need to focus on is that these competing conceptualizations and periodizations yield alternative ways of seeing the Western philosophical tradition: as infected by racism only from modernity onwards, or as possibly infected by racism from the start. As for the why: the answers given will obviously be shaped by the answers one finds plausible to the questions about conceptualization, periodization, and place of birth (the what, when, and where). Among the wide range of proposed origins and theories offered about those origins are: universal human ethnocentrism mutating into a racialized form; religio-culturalist biases; European color symbolism (for white racism in particular); unconscious psychosexual dynamics; rational-choice power politics; sociobiology; cognitive psychology (perhaps as a theoretical complement to the ethnocentric account); and Marxist political economy.

3. The History of Philosophy

Histories of philosophy standardly cover a number of themes: the detailed study of the work and ideas of individual philosophers in relation to their times, intellectual ancestors, and contemporaneous friends and foes; the tracing of lines of filiation, influence, and critique; the exploration of the evolution and development of specific concepts and theoretical frameworks over the centuries; the identification of schools of thought; the demarcation

and highlighting of particular periods and geographical areas as exemplifying and locating illuminating intellectual trends (the "German Enlightenment"); and so forth.

How would an interest in race and racism orient one's approach to the history of philosophy? One obvious line of research would be to try to track the origination and development of race and racism as concepts and framing presuppositions in the work of Western philosophers, and to excavate the ways in which such assumptions might have shaped their thought (West 1982; Goldberg 1993; Eze 1997b; Mills 1997; Bernasconi 2001; Bernasconi and Lott 2000; Ward and Lott 2002; Valls 2005b; Eigen and Larrimore 2006; Sala-Molins 2006). On this basis, one would then seek to construct a revisionist history of the discipline that acknowledged and explored the role of race-thinking across philosophy's various fields (metaphysics, epistemology, value theory, social and political philosophy, existentialism and phenomenology, etc.), and raised the question of how conventional aracial narratives and frameworks might then need to be reconceived. In addition, anti-racist philosophizing by recognized Western philosophers, as well as by those thinkers not currently recognized as philosophers (Du Bois, Fanon), could then be readily integrated into a narrative no longer sanitized to be acceptable to current norms. Uncritical philosophy of race and critical philosophy of race would then be part of the *same* story, rather than the latter's being represented as a perverse shadow-boxing, nothing to do with the canonical figures' work, or indeed with philosophy at all as properly understood.

3.1 The Pre-Modern Period

On the revisionist periodization of Delacampagne, Isaac et al., the appropriate starting point is with the thinkers of Greco-Roman Western antiquity, beginning with the Athens of the fifth century BCE. Isaac emphasizes that their conception of race is different from the modern one: not necessarily color-coded and not biologically determinist. Nonetheless, they are for him conceptions of *race* and *racism* ("proto"-racism originally), not just ethnocentrism and xenophobia. The crucial criterion, he suggests, is the positing of an innate differentiation, "physical and mental or moral," that grounds a hierarchy of superiors and inferiors which "cannot be changed by human will." Isaac points out that in the intellectual framework of the times, environmentalism was not sharply counterposed to hereditarianism as it is for us, since, in a kind of ur-Lamarckianism, human traits determined by climate or geography or social institutions could, it was believed, be passed on to subsequent generations (Isaac 2004, 34–38). He sees *Airs, Waters, Places*, traditionally attributed to Hippocrates, as having had "an enormous influence, not only on ancient philosophers such as Plato and Aristotle … but also on early modern authors" like Bodin, Montesquieu, Hume, and Herder (60): "Climate, geography and institutions all go together in producing peoples of uniformly good or bad character" (65).

For Isaac, it is above all Aristotle's views in the *Ethics* and the *Politics* on barbarians, the differences between Greeks and non-Greeks, and, most importantly, on "natural" slaves, that make him a, or the, pioneering racist philosopher of antiquity who needs to be recognized as such (172–81). Though the *Politics* may be non-committal on possible bodily markers of natural slaves, there are, in Isaac's opinion, enough passages in other writings to make it clear that "slaves by nature are non-Greeks and the masters by nature Greeks,

which means that the division between superior and inferior men is essentially one based on ethnic identity" (178). On this basis, Isaac argues that Aristotle should indeed be seen as a philosophical pioneer of racist theory—"[his] theory combines an imperialist ideology with an uncompromising proto-racist attitude" (181)—and that other thinkers from the Greco-Roman period influenced by *Airs* and Aristotle's views on natural slaves (including Tacitus, Cicero, and Seneca) should be categorized likewise. Moreover, insofar as the rediscovery in the Renaissance of the texts of classical antiquity is standardly seen as crucial to the rebirth of Western philosophy in the early modern period, the racist dimensions of these texts need to be highlighted and acknowledged as part of the classical heritage that is likewise passed down. Suitably updated for the world of the Scientific Revolution and the European conquest of the rest of the planet, they eventually come to serve as the template for racist theorizing in the contemporary form we would now recognize.

What, however, about the role of religion, and the at least nominally universalist creed of Christianity, in the hundreds of intervening years of the Middle Ages? The problem is that in Christian anti-Semitism we have ample precedent for the framing of a people as beyond Christian redemption and possibly not even human. Even Fredrickson (2002), for whom racism is modern, concedes that "the popular [medieval] belief that all Jews were in league with the Devil scarcely encouraged a firm conviction that they were fellow human beings" (21). Although biological ("scientific") racism was the dominant variety from the late eighteenth century to the mid-twentieth century, "cultural" racism has been argued by some scholars to be historically the modal form, with biological racism the outlier. On this analysis, at least some varieties of medieval anti-Semitism and Islamophobia should legitimately count as racism, so that the Spanish *limpieza de sangre* criterion would not (contra Fredrickson) signify the shift from religious anti-Judaism to racial anti-Semitism but the shift from *culturally racist* anti-Semitism to *biologically racist* anti-Semitism.

Moreover, the "monstrous races" bequeathed to medieval Christendom from Pliny the Elder's *Natural History*—such as the headless (Blemmyae), the one-legged (Sciopods), and the dog-headed (Cynocephali)—would furnish an iconography that could be developed and extended into a bestiary of unbelievers: human enemies of the faith who turned out to be creatures equally bizarre and threatening. In Debra Strickland's study of medieval art, *Saracens, Demons, and Jews* (2003), she suggests that "[I]n the medieval imagination, Monstrous Races, black Africans, Jews, Muslims, and Mongols share many of the same physical, moral, and behavioral characteristics," having "deformed bodies, strange dwellings, barbaric habits, and sinful behaviors" (59). Similarly, some scholars have argued that in antiquity the Ethiopian was seen as intrinsically evil because of his color, and that this negative perception of blacks is inherited by Christianity and perpetuated in images of black demons torturing Christ, and depictions of the Devil himself as black (Goldenberg 2009).

Where did these problematic humans come from? Polygenesis was, of course, a heretical doctrine for Christians. But radically differentiated traits and characteristics could be reconciled with monogenesis through the story in the Hebrew Bible of the Curse of Ham (or sometimes the Curse upon Cain), which was applied to blacks in particular (Goldenberg 2003). Because of Ham's allegedly staring at Noah in his nakedness after the Ark's descent on Mount Ararat (though some interpretations suggest sexual impropriety as well), Ham's son Canaan, was cursed by Noah to have all his descendants be slaves. According to later

readings of the biblical "Table of Nations," the three sons then went their separate ways to found the (then-known) three continental populations of Europe (Japheth), Asia (Shem), and Africa (Ham). So though the passage from Genesis (9: 18–27) did not itself mention skin color, this became a popular explanation in the Arab world, later transmitted to the Christian world, for natural black servility. An extensive medieval Arab slave trade long predated the Atlantic slave trade, and while people from all ethnic groups were enslaved, evidence suggests that blacks were typically given the most menial and degrading jobs. In an essay on medieval Muslim philosophers on race, Paul-A. Hardy (2002) points out that the Arab word for blacks, *abīd*, is virtually synonymous with "slaves" (as "Negroes" would come to be later for the Atlantic community). Such well-known figures as Abu Ali Ibn Sina (Avicenna) and Ibn Khaldun asserted that blacks (along with Slavs and Turks, the preferred communities for Muslim enslavement) were natural slaves, and in the Jewish philosopher Maimonides' *Guide to the Perplexed*, he opined that blacks were not on the level of human beings, but somewhere in between monkeys and the human, a trope that would of course reappear with great impact in modern racism. Whether or not we want to term these views racist, they are obviously important for the (pre?)history of the concept.

Finally, apart from such sources, contributors to the *Origins of Racism in the West* (2009) conference volume argue variously that Christian communities were explicitly or implicitly ethnically differentiated and that a medieval science was developing that was indeed (in keeping with the "Physiognomics" inherited from the ancient world) capable of making the proto-biological distinctions among humans usually relegated to the period several hundred years later.

3.2 The Modern Period

However, it is obviously with modernity that race uncontroversially becomes a crucial social category. The term itself appears in various European languages over the fifteenth–sixteenth centuries, used as a collective noun to denote not just people but entities of all kinds. Gradually, however, it becomes restricted to human beings. Various theorists—Bernier, Linnaeus, Buffon, Blumenbach, and others—have been given the (dis)credit for developing the concept of race, or (on the race as pre-modern periodization), the modern concept of race. But our concern here is with philosophers in particular.

Consider a standard line-up of important modern philosophers: Hobbes, Locke, Hume, Rousseau, Kant, Hegel, Mill, Marx. We can ask such questions as: Did they express racist views? Were these views theorized or offhand remarks? Were they incidental or central to their thought? Were they important to the development of the concept of race, and the theory of racism? Have they come to shape modern philosophical thought, and if so, how? Necessarily, my comments here will be very brief, and should be taken as indicative rather than definitive, as an incentive for readers to investigate the growing secondary literature on these questions for themselves.

Hobbes's exclusion of Native Americans from the ranks of those humans rational enough to exit the state of nature has been seen as racist by some commentators, but not others. Though reference to Amerindians occupies only a few sentences in his key texts (*Leviathan, On the Citizen*), some scholars (Ashcraft 1972; Hall 2005) have argued that the

conceptualization is nonetheless theoretically important in legitimizing the category of the apolitical "savage," not part of civil society, which would play such a crucial role in Western colonial thought. Locke's representation of Amerindians in the *Second Treatise* has likewise been seen as justifying the colonial project, especially his judgment in ch. 5 that "In the beginning all the world was America," that is, uncultivated land which required the hand of European industry to "mix itself with nature" (Tully 1993; Arneil 1996). His earlier investments in African slavery and later colonial administrative role, and assistance in authoring the Carolina Constitution, which gave masters despotic power over their slaves (a slavery which was hereditary), have also been cited as evidence that blacks were racially excluded from equal status with the "men" who in the *Second Treatise* have natural rights to freedom (Bernasconi and Mann 2005). However, others have argued in reply that although Locke's hands were undoubtedly "dirty," commonplace human hypocrisy and inconsistency—"a kink in his head"—are to blame rather than any systematized theoretical views on black inferiority (Farr 2008). Hume's notorious footnote in his essay "Of National Characters" asserts that only whites have created a civilized nation; again, some scholars (Valls 2005a) have argued that—while of course deplorable—it is an offhand remark to which we should not attribute much philosophical significance. Rousseau is an ambiguous figure. On the one hand, he sketches in *Discourse on Inequality* an environmentalist–technological account of evolutionary human development and degeneration that many commentators have seen as anticipating (or even inspiring) Marxism. The "savage" man is physically and psychologically healthier than the degraded "civilized" man of a society based on private property and self-seeking wills, his amour-de-soi not having degraded into amour-propre (Moran III 2002). On the other hand, Rousseau's real-life examples of "savages" ("natural men") are limited to people of color (Boxill 2005), and the contemporaneous case of New World slavery, such as the French colony of Saint Domingue (Haiti), is oddly absent from his seemingly unequivocal condemnation of slavery in the opening chapters of the *Social Contract*, leading some to argue that his environmentalism becomes qualified where blacks are concerned (Sala-Molins 2006).

Kant has been the subject of particular interest, both because of his importance to modern philosophy and the Enlightenment, and the overtly racist anthropological and geographical writings which have led some commentators to bestow on him the dishonor of being the father, or at least a father, of modern "scientific" racism (Eze 1997a; Bernasconi 2002). Debate has been centered on the question of whether Kant changed his mind (a racist Kant succeeded by an anti-racist Kant) (Muthu 2003; Kleingeld 2007; Bernasconi 2011), and, whether he did or not, what implications the passages in the anthropology and physical geography should be seen to have for his metaphysics, epistemology, teleology, aesthetics, and moral and political theory. Should they be seen as embarrassing but theoretically irrelevant manifestations of bigotry, which can be purged without affecting the theory in question (the dominant position of his defenders) (Louden 2000; Hill and Boxill 2001)? Or should they be seen as shaping his views to a greater or lesser degree, for example in claims about who competent knowers can be, which race will have the vanguard role in leading humanity to the global kingdom of ends and which races might not make it at all, whether all humans do attain the status of moral equality and if not, what the implications are for the structure of the *Rechtsstaat* (Eze 1997a; Bernasconi 2002, 2011; Mills 2005)?

John Stuart Mill's *On Liberty*, a staple of thousands of liberal arts courses over the past few decades, stipulates that his anti-paternalist harm principle is not meant to apply to races in their "nonage," barbarians for whom despotism is legitimate—a passage seemingly detached from and incongruent with the rest of the book if students have not been given the background information that Mill was also a colonial theorist, an employee of the British East India Company. Similarly, *Considerations on Representative Government* makes plain why Indians are not yet mature enough to govern themselves (Mehta 1999; Pitts 2005). Mill has a famous exchange with Thomas Carlyle about the post-Emancipation state of West Indian blacks, but though Mill's is the progressive position (compared with the biologically racist Carlyle), this is only relative. He thinks that blacks are culturally backward and in need of European enlightenment, though admittedly making reference to Ancient Egypt as a black civilization, as contemporary Afrocentrists do (Goldberg 2002a; Bogues 2005).

Hegel's philosophy of history is influenced by a geographical determinism which denies historical achievement to Africans and Native Americans, and denominates Europe and Europeans as the truly world-historical continent and peoples (Tibebu 2011). Marx and Engels's historical materialism would seem to be intrinsically anti-racist, given its social environmentalist assumptions. But insofar as they inherit a Hegelian view of world-historical and non-world-historical peoples (which includes various European ethnic communities), albeit materialistically rather than idealistically understood, transformative global agency for Marx and Engels tends to inhere in the white proletariat. Moreover, their view of "Asiatic despotism," an ad hoc category awkwardly fitting the Euro-timeline of primitive communism, slavery, feudalism, capitalism, and advanced communism, which contrasts a static East with a dynamic West, has been indicted by postcolonial theorists as Eurocentric, if not racist (Hobson 2012).

The recovery and reconstruction of an oppositional strain of anti-racist thought among European philosophers is also an important task that needs more work. (The analogy here would be with John Stuart Mill as a pioneering male feminist.) Hume's comments on races' divergent civilizational capacities were opposed by James Beattie; Herder argued with Kant over his concept of race; Burke condemned British injustices in India. A problem, however—and this applies to Mill as feminist also—is that even (in some respects) anti-racist theorists continued to espouse views that (in other respects) might now be seen as racist. Burke did not condemn empire itself. Montesquieu satirized the racism that accompanied some justifications for African slavery but did not actually call for Negro emancipation. And in general, it needs to be borne in mind that even actual abolitionism and consistent anti-imperialism could—and often did—coexist with the belief in non-white inferiority.

Three important twentieth-century European figures for anti-racist theorization are Hannah Arendt, Jean-Paul Sartre, and Michel Foucault. Arendt's (1973) book on totalitarianism has achieved increasing celebrity in recent years as a statement of the connection between European imperialism and the Holocaust, the "boomerang thesis" that links Nazism to an earlier history of modernity and genocide. But some critics have found her periodization of racism problematic, and her own comments on Africans themselves racist (King 2004; Gines 2014). Sartre's preface to Fanon's *Wretched of the Earth* (2004) is deservedly famous, but he has many other lesser-known anti-colonial writings also (Sartre 2006) that, for a growing number of commentators, earn him a place as one of the key

twentieth-century philosophical theorizers of colonialism and racism. Foucault's analyses of state racism and "biopower," for example in *"Society Must Be Defended"* (2003), have been found both fruitful for understanding racial domination and frustrating for their idiosyncratically generic conception of "race," that marginalizes the experience of blacks under New World slavery and people of color under European colonialism (despite France's central role in both), and frames Stalinism as a major twentieth-century form of "racism."[3]

4. Contemporary Critical Philosophy of Race

We turn now to contemporary critical philosophy of race as it is being developed by professional philosophers in the standard areas recognized by the analytic and Continental traditions. What difference does race make (if it does) to the way we do metaphysics, epistemology, value theory (aesthetics, ethics), social and political philosophy, and so forth, or phenomenology, existentialism, and pragmatism? (Some of the ways it would affect the history of philosophy have been covered above.) Obviously, all I can provide here is a brief overview that I hope will stimulate interested readers to undertake more detailed explorations of their own. Two valuable introductory texts are Atkin (2012) and Taylor (2013).

4. Metaphysics

Some feminist theorists have argued that divergent gender socialization and experience run so deep that they shape one's metaphysical picture of the world even in areas with no apparent connection with gender ("androcentric versus gynocentric metaphysics"), such as whether the everyday world is thought of in terms of discrete particulars or interdependent and interrelating entities (Haslanger 2000, 108–9). But the divergent metaphysics of, say, Native American, African, Asian, and Australian cosmologies—for example, in views on causality and humanity's place in the universe—pre-existed European racial subordination, and as such would not be conceptually equivalent to gender-based dimorphic ontologies. However, a case could be made that these non-Western metaphysical systems (presumptively ethnic and national in origin rather than "racial") might subsequently *acquire* an oppositional racial dimension to the extent that they are conscripted into and reshaped by the ideological struggle against white racial domination. Black liberation theology, for example, can be found in a variety of syncretized religious practices across the diaspora in the United States, Latin America, and the Caribbean—the black American church, Brazilian *candomblé*, Cuban *santería*, Haitian *vodun*, Jamaican Revival Zion—that have fused African religions and elements of Christianity into a source of spiritual sustenance and militancy against, first, slavery and then post-emancipation racial subordination (Murphy 1995).

[3] Thanks to Jerry Miller for guidance on this point.

The more straightforward and uncontroversial examples are the attempts to develop a metaphysics of race itself, both for society (races as social existents—and if so of what kind—or non-existents) and the self (selves as racial or aracial). The early 1980s'–1990s' phase of critical philosophy of race as a self-conscious enterprise was marked by debates over the question of whether, if race did not exist biologically, it could be said to exist at all. W. E. B. Du Bois's classic 1897 essay, "The Conservation of Races" (Sundquist 1996), was a frequent reference-point, insofar as Du Bois, in arguing for the need to "conserve" race and race-based organizations as part of an emancipatory African American political project, used ambiguous language that made it unclear whether he thought of races as natural or social. (Chike Jeffers [2013] has recently pointed out a further ambiguity in the text between socio-political and socio-culturalist racial conceptions, and contends that Du Bois endorsed the latter.)

Three seemingly clearly demarcated positions could be said to have coalesced out of these debates: racial naturalism, a minority position (races do indeed exist biologically, though not in the racist hierarchy traditionally supposed); racial eliminativism or skepticism, another minority position (race has neither a natural nor a social existence); and anti-eliminativist racial constructionism, the majority position (race does exist, but as a social construct—not a natural kind but a social kind). Anthony Appiah (1992) and Naomi Zack (2001) were the two figures most prominently associated with an eliminativist position, drawing on theories of reference from philosophy of language as well as the findings of biology, genetics, and anthropology to argue that races did not exist. However, most other critical philosophers of race endorsed anti-eliminativism (e.g. Gordon 1995, 1997b, 2000; Outlaw 1996, 2005; Mills 1997, 1998; Gracia 2005; Stubblefield 2005; Gooding-Williams 2005, 2009; Alcoff 2006; Sundstrom 2008; Haslanger 2012; Taylor 2013), presuming that a social constructionist historical account of race could be given that avoided Appiah's objection that such claims ran into problems of circularity. (Appiah [1992, 32] had argued that "sharing a common group history cannot be a *criterion* for being members of the same group, for we would have to be able to identify the group in order to identify *its* history.")

Even at the time, this ideally sharp three-way partitioning was in actuality muddied by various qualifying claims. Appiah (Appiah and Gutmann 1998) conceded, for example, that even if races did not exist, racial identities could. Outlaw (1996) did not actually support orthodox constructionism but a hybrid "social-natural" position, according to which races were *both* social and natural, insofar as they were formed out of "bio-cultural" endogamous reproductive patterns. Bernard Boxill (2001a) raised the possibility that while the races as we now know them might be social constructs, this did not rule out there being biological races as well. Anna Stubblefield (2005) recommended we think of races as families, though this recommendation was meant more prescriptively than descriptively. Jorge Gracia outlined both a "Genetic Common-Bundle View" of race (2005, ch. 4) and a "Familial-Historical View" (2007, ch. 3).

Since then the debate has become vastly more complicated, increasingly relying on technical distinctions and concepts from the philosophy of language (neo-descriptivism versus reference externalism). What had seemed to be clear-cut lines of division have been breached with the synthesizing of various hybrid positions that draw selectively from different strains of naturalism, eliminativism, and constructionism. Ron Mallon (2012), for example, suggests that in actuality "a univocal constructionist account of race" is

unattainable, that "skepticism and the varieties of constructionism share a broad base of metaphysical agreement" (55), and that the aim should be to develop an account geared to our varying "functional needs" (76). Robin Andreasen (2012) and Philip Kitcher (2012) argue independently that the constructionist consensus is mistaken insofar as it asserts the biological unreality of race, and that races do exist, whether as cladistic subspecies (Andreasen) or inbred lineages from reproductive isolation (Kitcher). But both agree that "these conceptions can be complementary" (Andreasen 2012, 83) and that, contra common sense, there is no contradiction here: "Races are both biologically real and socially constructed" (Kitcher 2012, 109). The appearance of paradox arises from our presupposing a strongly realist and essentialist view of natural kinds, which Kitcher believes we should abandon in favor of a weaker "pragmatist" account that recognizes "the pluralistic character of taxonomi[c] practices in the sciences, especially within biology." We are not really carving nature at its joints but developing a multiplicity of frameworks adapted for different kinds of inquiry (110). On the other hand, Joshua Glasgow (2009b) rejects all three of the positions as originally understood: naturalism and biological race, the framing of constructionism as a realist position, and traditional eliminativism. He advocates what he calls "racial reconstructionism," a substitutionist rather than eliminativist position, for which classic racial discourse needs to be replaced with a "proximate" post-reconstruction discourse in which "race*"—a completely social kind—is completely stripped of any biological link such as "race" had. Sally Haslanger (2012, 429–45) seeks to develop a "rational improvisation" approach that rejects both "a priori intuitions" and "[linguistic] experiments concerning the 'folk theory' of race" (440). Alternative metaphysical positions will doubtless continue to proliferate (*The Monist* 2010).

4.2 Epistemology

The recent development of social epistemology as a recognized area within the field has provided a respectable entry-point for socially informed philosophical investigations into group-based cognitive processes and norms, including politically radical perspectives (Marxist, feminist), that were long excluded or marginalized by the orthodox Cartesian individualist framework. For the left and feminist traditions, of course, society is explicitly conceived of in terms of group domination and subordination (classes, genders), so that the consequences of social oppression for social cognition are placed at epistemological center-stage. Unfortunately, philosophical work on race has lagged behind this growing literature. But white racial domination has obviously been similarly powerful in negatively influencing social cognition, at least for the past few hundred years, so that the exploration of the processes by which it worked (works) and the norms by which it needs to be resisted would presumably be similarly theoretically rewarding.

Across a variety of other disciplines, scholars have sought to tease out the constitutive elements of "whiteness" as a cognitive framework, so this project is already well under way elsewhere. (See, for example, sociologist Joe Feagin's [2010] concept of the "white racial frame.") The task of philosophy and formal epistemology would then be to take this investigation to the higher level of abstraction and generality required by disciplinary norms. David Theo Goldberg's (1993) and Ladelle McWhorter's (2009) use of Foucauldian

discourse theory, Lewis Gordon's (1995) drawing on Sartrean "bad faith" to elucidate the dynamic of racialized cognition, Linda Martín Alcoff's (2006) adoption of Hans-Georg Gadamer to map out how social identities establish interpretive horizons for us, Marilyn Frye's (2001) proposal of "whiteliness" as a concept for capturing the distinctive optic of white racial privilege, Charles Mills's (1997) positing of a white "epistemology of ignorance" that ensures a self-reproducing, strategically functional "white ignorance" are all attempts, from different theoretical perspectives, Continental and analytic, to get at the processes of racially biased (structural or motivated) non-knowing—the formalization of Ralph Ellison's (1995) novelistic insights, in *Invisible Man*, about the "peculiar disposition" of "the construction of [whites'] *inner* eyes" (3). Race and the epistemology of ignorance has been the subject of a conference volume (Sullivan and Tuana 2007). In analytic circles, a growing body of work is making use of the implicit bias literature from psychology, which has revealed the alarming extent to which even people who sincerely think of themselves as liberal continue to have racist and sexist preconceptions.

Racial versions of "standpoint theory" are also being developed, extrapolating on Du Bois's insights about the "second sight" produced by racial subordination (Gooding-Williams 2009, ch. 2). "Intersectionality"—the investigation of the multifaceted experiential dimensions of having multiple interacting identities (of gender, race, sexual orientation, etc.)—which has established such a presence in other disciplines, is virtually non-existent in philosophy, where women of color do not number more than a few dozen (Zack 2000). But Anika Maaza Mann (2010) has argued for an intersectional standpoint theory, Ladelle McWhorter (2009) provides an integrated Foucauldian account of racism and sexual oppression, and one of George Yancy's (2010) many edited collections brings together white women in self-conscious and self-critical examination, from their gender-disadvantaged but racially privileged perspective, of the whiteness of philosophy. José Medina's (2012) book-length treatment of the "epistemology of resistance," that self-consciously mines queer theory and the feminism of women of color as well as critical race theory for insights, is probably the most developed articulation of such an epistemology to date.

4.3 Value Theory

4.3.1 Aesthetics

Aesthetics is another underdeveloped area, though discussions about the way in which race affects structures of feelings, judgments on the beautiful and ugly, and the theorization of the aesthetic itself go back hundreds of years. The resistance of people of color to racial subordination has perforce had an aesthetic dimension: the challenge to racist norms, the transgressive re-incorporation of the body, the imperative of learning to *feel* in a different way, the critique of existing works of art, and the production of new work informed by different values. For the past century, for example, all the major political movements of the black diaspora have had as part of the agenda the construction of a *counter*-aesthetic—the "New Negro" and the Harlem Renaissance, Negritude and the cultural decolonization of the anti-colonial struggle, the Black Power and the Black Arts Movement of the 1960s.

Again, however, self-conscious theorization within contemporary philosophical circles of these issues has been rare, despite the precedent set by Alain Locke, though this is beginning to change.

Tommy Lott's (1999) collected cultural essays on black cinema, representations of blacks on television, black youth culture, rap, the Negro–ape metaphor in racist discourse, and the black painter Sargent Johnson were one early attempt to theorize the aesthetics of race. The most famous woman of color with a philosophy degree, Angela Davis (1999), published the same year a study of the blues and black feminism which also has obvious relevance to this theme, if not self-consciously explored as such. Since then, Derrick Darby and Tommie Shelby (2005) have edited a collection on philosophy and hip-hop for the Open Court Philosophy and Popular Culture series, and Dan Flory (2008) has produced a full-length book on the subject of a distinctively African American variety of film noir, a noir doubly black, so to speak. A special volume of the online journal *Contemporary Aesthetics* (2009) on aesthetics and race edited by Monique Roelofs discusses such issues as mixed-race looks, taste and white embodiment, Garveyist aesthetics, Bollywood cinema, and Hurricane Katrina as racial spectacle. Occasional pieces on film, plays, and literature have also been done by Lewis Gordon (1997b), Robert Gooding-Williams (2005), and George Yancy (2008, 2012), among others, and Flory and Mary Bloodsworth-Lugo (2013) have a recent edited collection on philosophy, race, and film, in which the authors focus on the overt or covert racial texts of films like *Twilight, Avatar, The Matrix*, and *The Help*. So perhaps this presages a trend to greater activity in the field, which would be welcome considering the obvious connections that can be made to the huge body of work already existing in cultural and ethnic studies.

4.3.2 Ethics

Far more work has been done in the other subdivision of value theory, ethics. In the analytic tradition, ethics is standardly divided into applied ethics, normative ethics, and meta-ethics. Applied ethics was open to certain kinds of discussions on race from the start, long before other areas of philosophy were, because of the debates from the 1970s onwards about affirmative action and preferential admissions. Whether all this literature should be seen as critical philosophy of race merely because it deals with race is an open question; I am doubtful myself. But at any rate, it is so large, familiar, and well represented in innumerable introductory philosophy textbooks that there seems no need to discuss it further.

More interesting and relevant is the normative analysis of racism, but since this topic overlaps with social philosophy, I will postpone it till the next section. The moral questions raised by race-based institutions like American slavery, and (arguably) their race-based legacy, such as the black "underclass," have been examined by various authors (Boxill 1992; McGary and Lawson 1992; Lawson 1992; McGary 1999; Blum 2002; Corlett 2003; *Journal of Ethics* 2003; Stubblefield 2005; Sundstrom 2008; *Journal of Social Philosophy* 2010). Naomi Zack (2011) has recently proposed that the problems of race are so serious and enduring that we need a distinctive racial ethic to tackle them. The moral implications of "whiteness" as social system, reproductive norm, state of mind, "gaze," "habit," and privileged social positioning have also been explored (Alcoff 2006; *Hypatia* 2007; Sullivan 2006; Yancy 2004, 2008, 2010, 2012). In a widely discussed essay, South African philosopher Samantha

Vice (2010) raised the question of whether whites in post-apartheid South Africa could lead virtuous lives, leading to a symposium in the *South African Journal of Philosophy* (2011). Meta-ethical issues are beginning to get some attention. In his *Racist Culture* (1993, ch. 2), David Theo Goldberg had pointed out the racialization of modern Western ethical systems, and Derrick Darby (2009) has argued that the history of the racial subordination of blacks is most illuminatingly to be thought of as showing not that blacks under white supremacy had moral rights that were not being respected, but that blacks did not have moral rights at all. All rights are the product of social recognition, and the natural rights tradition is wrong.

4.4 Social and Political Philosophy, Descriptive and Normative

As one would expect, however, it is in social and political philosophy that the most extensive work on critical race philosophy has been done. The workings of race at the macro and micro levels, and the normative consequences for theory and practice, have generated a significant body of literature.

To begin with, competing accounts have been offered of racism itself, whether as doxastic/cognitive, attitudinal/volitional, or behavioral, and its relation to social structure. In an important early formulation, Kwame Anthony Appiah (2012) contrasted extrinsic racism (belief in racial superiority/inferiority resting on putatively factual claims about the "races") with intrinsic racism (belief in racial superiority/inferiority merely through racial membership, independent of such claims). Jorge Garcia, in a series of influential papers, beginning with "The Heart of Racism" (2012), argued that belief is non-essential to racism—neither necessary nor sufficient—since the crucial factor is really ill will towards people because of their race, from outright hatred to simple indifference. Various critics—Tommie Shelby (2012), Levine and Pataki (2004)—have demurred. Joshua Glasgow (2009a) makes a case for racism as race-based disrespect, thereby in his mind solving the "location problem" of where to situate racism (beliefs or attitudes or practices). Worried about the coarseness of the categories of contemporary discourse, Lawrence Blum (2002) offers a more subtilized and nuanced range of terms, such as "racial insensitivity" and "racial ignorance," that people may be guilty of without being "racist." Important discussions can also be found in Harris (1999) and Boxill (2001b).

The relation between race, individual racism, and broader institutions and social structures has also been conceptualized in different ways that are, unsurprisingly, linked to the philosopher's larger socio-political commitments: racism as a Foucauldian discourse (Goldberg 1993; McWhorter 2009); racism as "bad faith" (Gordon 1995); racism as an individual vice that then infects social institutions (Garcia 2012); race and racism as white "contractarian" constructs (Mills 1997); racism as an ideology institutionally generated to rationalize exploitation (Shelby 2012); race as a Heideggerian "technology" used by the state to control "unruly" populations (Sheth 2009).

But race also provides a framework for the rethinking of the orthodox categories, ideologies, and concerns of mainstream social and political theory. As cited above, slavery, the "underclass," "whiteness," and "white supremacy" have all been taken as appropriate

theoretical objects for philosophical investigation. Tommie Shelby (2005) analyzes and prescribes what he sees as the appropriate philosophical foundations of black solidarity, and Robert Gooding-Williams (2009) compares the political visions and organizational strategies of W. E. B. Du Bois and Frederick Douglass as examples of "Afro-modern political thought" aimed at overcoming white supremacy. Racial liberalism (*Southern Journal of Philosophy* 2009) and racial nationalism (Ortega and Alcoff 2009) have been surveyed. Moreover, if the nation-state has been the standard unit of analysis for modern Western political theory and political philosophy, the anti-imperialist tradition of people of color, such as Du Bois (Sundquist 1996) and Fanon (2004), has often theoretically embedded local struggles in an awareness of a broader global macro-polity, the world of European colonialism and imperialism. Racial domination should be seen as planetary, liberalism as (for the most part) imperial (Pitts 2005; Hobson 2012), and the modern state in general as racialized (Goldberg 1993, 2002b; Mills 1997). Olufemi Taiwo (2010) looks at the processes by which colonialism retarded Africa's moves toward modernity. Such work has implications, not just for the history of philosophy, but for the philosophy of history. Thomas McCarthy (2009) proposes a new Kant-inspired philosophy of history informed both by awareness of Kant's racism and the theoretical resources Kantianism (duly sanitized of course) provides for a moral imperative of planetary transformation.

And that brings us naturally to the issue of racial justice, both local and global. For the past four decades, since the work of John Rawls, the central theme of Anglo-American political philosophy has been shifted from the issue of political obligation to the question of the justice of a society's "basic structure." But whether dealing with national or global justice, the issue of race and racial justice is almost never mentioned. An exception is Tommie Shelby (2004), who has argued that Rawls's fair equality of opportunity (FEO) principle can be developed and extended to remedy the legacy of racial discrimination. But other political philosophers working in the liberal tradition have been more skeptical (Boxill 1992; McGary 1999; Pateman and Mills 2007; Graham 2010). Elizabeth Anderson (2010) explicitly repudiates Rawls's ideal theory as methodologically incapable of providing the appropriate guidance for solving the non-ideal problem of race. Rodney Roberts (2002) points to the relative under-theorization of rectificatory justice in the literature as a manifestation of social privilege. Here Daniel Butt's (2009) recent case for inter-continental rectificatory justice is a welcome contribution that could arguably be supplemented with a racial analysis.

But mainstream liberalism, of course, has never been the only political game in town. In the long nineteenth- and twentieth-century struggle of people of color against colonialism, imperialism, and national and transnational white domination, many theorists were attracted to more radical ideologies: Marxism and ethnoracial nationalisms of one variety or another (black nationalism, Pan Africanism), and attempted syntheses of the two—"black" and Third World Marxisms. In Mexico, José Vasconcelos's (1997) concept of a "cosmic race" produced by racial mixture, *mestizaje*, played a role in the development of *indigenismo* as a national ideology. Lucius Outlaw's pioneering work on critical philosophy of race (1996, 2005) was located in part in the "critical theory" tradition that ultimately derives from Marx, though critical of it for its Eurocentrism and failure to pay attention to white supremacy. Cornel West (1982, 1989) tried to synthesize black liberation theology, Marxism, and Deweyan pragmatism into a racially-informed emancipatory ideology he

called "prophetic pragmatism," and Bill Lawson and Donald Koch (2004) co-edited a collection on pragmatism and the problem of race. Tommy Curry (2011), in a series of forceful essays, argues (following the late legal theorist Derrick Bell) that racism in the United States is permanent, indicts what he sees as the "integrationist agenda" of dominant figures in African American philosophy, and calls instead for a "culturalogical" working out of the diasporic black tradition's political prescriptions.

Finally, while as mentioned blacks have understandably been differentially represented among philosophers of color in critical philosophy of race, recent work has warned of the dangers of what has come to be called "the black/white paradigm" or "the black/white binary" for understanding race (Alcoff 2006; Sundstrom 2008; *Critical Philosophy of Race* 2013). Anti-Native American, anti-Asian, and anti-Latino racism have distinctive features of their own that cannot be simply assimilated to anti-black racism. Moreover, if critical philosophy of race is to be successful as a global enterprise it will obviously require sensitivity to local racial dynamics rather than the attempt to force them all into a US model. Falguni Sheth (2009) has tried to broaden the canvas by looking at the distinctive patterns of racialization of Muslim and Middle Eastern populations in the United States, and the way their racial status has changed after the 9/11 terrorist attack. The complications of categorizing Latinos/Hispanics—are they an ethnic group composed of many races, a race themselves, or both?—have also been examined (Gracia 2007).

4.5 Existentialism, Phenomenology, and Pragmatism

Though the hegemony of analytic philosophy in the Anglo-American world makes the achievement of respectability by analytic standards the bar of acceptability for critical philosophy of race, a case can easily be made that the Continental tradition (Bernasconi and Cook 2003; Gines 2012) has been far more important for its actual development. In his overview from the phenomenological perspective, Robert Bernasconi (2012) points out the influence of Edmund Husserl, and, more importantly, Jean-Paul Sartre and Maurice Merleau-Ponty, on Frantz Fanon's *Black Skin, White Masks* (2008), one of the key texts of the field. The distinctive existential agonies of blackness in a white, anti-black world are not captured by the *p*'s and *q*'s of logical analysis, and many philosophers have found this apparatus inadequate or alienating for dealing with the "lived experience" of race. In Du Bois's "double consciousness," Richard Wright's "underground man," Ralph Ellison's "invisibility," and Fanon's "zone of non-being," an existential phenomenology has been charted that, as Lewis Gordon (1997a, 2000) emphasizes, must not be assimilated to the Euro-existentialism of Kierkegaard and Camus, even if they both fall under the broader genus of what Gordon calls "philosophy of existence." Here the "absurdity" and "tragedy" of life arise less from the revelation that God is dead, or the finitude of human existence in an indifferent universe, than the day-to-day problems of negotiating a world shaped by a white domination aspiring to omnipotence and ensuring black existence falls below the human. Gordon's (1995, 1997b, 2000) own work has been central to exploring the peculiar features of this racialized existence, as has Linda Martín Alcoff's (2006) reconstruction of the phenomenology of the racialized body. "Mixed-race" status in its peculiar social, metaphysical, and existential aspects has been examined sympathetically by Naomi Zack (1993) in the

US context, and by Alcoff (2006, ch. 12) with respect to Latin *mestizaje*, and more critically by Gordon (1997b, ch. 3). From a pragmatist perspective, Shannon Sullivan (2006) draws on Deweyan "habit" to reconstruct the corporeal logic of white privilege, while George Yancy has undertaken a detailed analysis both of the black body under the white gaze (2008) and, turning things around, the white body and whiteness subjected to the black gaze (2012).

As mentioned, women of color are even more under-represented in the field than males (Zack 2000), but of the small number there are, many have found Continental philosophy more useful than analytic philosophy in tracing the intersectional issues of race and gender. *Convergences* (Davidson et al. 2010) looks at the links between black feminism and Continental philosophy, and *Living Alterities*, a collection with mostly female contributors edited by Emily Lee (2014), explores phenomenology, embodiment, and race.

References

Alcoff, Linda Martín. 2006. *Visible Identities: Race, Gender, and the Self*. New York: Oxford University Press.

Anderson, Elizabeth. 2010. *The Imperative of Integration*. Princeton, NJ: Princeton University Press.

Andreasen, Robin O. 2012 [1998]. "A New Perspective on the Race Debate." In Taylor (2012), vol. II, pp. 82–105.

Appiah, Kwame Anthony. 1992. *In My Father's House: Africa in the Philosophy of Culture*. New York: Oxford University Press.

Appiah, Kwame Anthony. 2012 [1990]. "Racisms." In Taylor (2012), vol. III, pp. 5–18.

Appiah, Kwame Anthony, and Amy Gutmann. 1998. *Color Conscious: The Political Morality of Race*. Princeton, NJ: Princeton University Press.

Arendt, Hannah. 1973 [1951]. *The Origins of Totalitarianism*. New York: Harcourt Brace Jovanovich.

Arneil, Barbara. 1996. *John Locke and America: The Defence of English Colonialism*. Oxford: Clarendon Press.

Ashcraft, Richard. 1972. "Leviathan Triumphant: Thomas Hobbes and the Politics of Wild Men." In Edward Dudley and Maximillian E. Novak, eds., *The Wild Man Within: An Image in Western Thought from the Renaissance to Romanticism*. Pittsburgh: University of Pittsburgh Press, pp. 141–81.

Atkin, Albert. 2012. *The Philosophy of Race*. Durham, UK: Acumen.

Bernasconi, Robert, ed. 2001. *Race*. Malden, MA: Blackwell.

Bernasconi, Robert. 2002. "Kant as an Unfamiliar Source of Racism." In Ward and Lott, pp. 145–66.

Bernasconi, Robert. 2011. "Kant's Third Thoughts on Race." In Stuart Elden and Eduardo Mendieta, eds. *Reading Kant's Geography*. Albany, NY: SUNY Press, pp. 291–318.

Bernasconi, Robert. 2012. "Critical Philosophy of Race." In Sebastian Luft and Søren Overgaard, eds. *The Routledge Companion to Phenomenology*. New York: Routledge, pp. 551–62.

Bernasconi, Robert, and Tommy L. Lott, eds. 2000. *The Idea of Race*. Indianapolis: Hackett.

Bernasconi, Robert, with Sybol Cook, eds. 2003. *Race and Racism in Continental Philosophy*. Bloomington: Indiana University Press.

Bernasconi, Robert, and Anika Maaza Mann. 2005. "The Contradictions of Racism: Locke, Slavery, and the *Two Treatises*." In Valls (2005b), pp. 89–107.

Bloodsworth-Lugo, Mary K., and Dan Flory, eds. 2013. *Race, Philosophy, and Film*. New York: Routledge.

Blum, Lawrence. 2002. *"I'm Not a Racist, But…." The Moral Quandary of Race*. Ithaca, NY: Cornell University Press.

Bogues, Anthony. 2005. "John Stuart Mill and 'The Negro Question': Race, Colonialism, and the Ladder of Civilization." In Valls (2005b), pp. 217–34.

Boxill, Bernard R. 1992. *Blacks and Social Justice*. Rev. ed. Orig. ed. 1984. Lanham, MD: Rowman & Littlefield.

Boxill, Bernard R. 2001a. Introduction to Boxill (2001b), pp. 1–42.

Boxill, Bernard R. ed. 2001b. *Race and Racism*. New York: Oxford University Press.

Boxill, Bernard R. 2005. "Rousseau, Natural Man, and Race." In Valls (2005b), pp. 150–68.

Butt, Daniel. 2009. *Rectifying International Injustice: Principles of Compensation and Restitution Between Nations*. New York: Oxford University Press.

Contemporary Aesthetics. 2009. Special Volume 2: Aesthetics and Race: New Philosophical Perspectives.

Corlett, Angelo. 2003. *Race, Racism, and Reparations*. Ithaca, NY: Cornell University Press.

Crenshaw, Kimberlé, Neil Gotanda, Gary Peller, and Kendall Thomas, eds. 1995. *Critical Race Theory: The Key Writings That Formed the Movement*. New York: The New Press.

Critical Philosophy of Race. 2013: 1, no. 1: Critical Philosophy of Race beyond the Black/White Binary: 28–124.

Curry, Tommy J. 2011. "On Derelict and Method: The Methodological Crisis of African-American Philosophy's Study of African-Descended People under an Integrationist Milieu." *Radical Philosophy Review* 14, no. 2: 139–64.

Darby, Derrick. 2009. *Rights, Race, and Recognition*. New York: Cambridge University Press.

Darby, Derrick, and Tommie Shelby, eds. 2005. *Hip-Hop and Philosophy: Rhyme 2 Reason*. Chicago: Open Court.

Davidson, Maria Del Guadalupe, Kathryn T. Gines, and Donna-Dale Marcano, eds., 2010. *Convergences: Black Feminism and Continental Philosophy*. Albany, NY: SUNY Press.

Davis, Angela Y. 1999. *Blues Legacies and Black Feminism: Gertrude "Ma" Rainey, Bessie Smith, and Billie Holiday*. New York: Vintage.

Delacampagne, Christian. 1983. *L'invention du racisme: Antiquité et Moyen-Age*. Paris: Fayard.

Eigen, Sara, and Mark Larrimore, eds. 2006. *The German Invention of Race*. Albany, NY: SUNY Press.

Eliav-Feldon, Miriam, Benjamin Isaac, and Joseph Ziegler, eds. 2009. *The Origins of Racism in the West*. New York: Cambridge University Press.

Ellison, Ralph. 1995 [1952]. *Invisible Man*. New York: Vintage.

Eze, Emanuel Chukwudi. 1997a. "The Color of Reason: The Idea of 'Race' in Kant's Anthropology." In Emmanuel Chukwudi Eze, ed. *Postcolonial African Philosophy: A Critical Reader*. Cambridge, MA: Blackwell, pp. 103–40.

Eze, Emanuel Chukwudi, ed. 1997b. *Race and the Enlightenment: A Reader*. Cambridge, MA: Blackwell.

Fanon, Frantz. 2004. *The Wretched of the Earth*. Trans. Richard Philcox. New York: Grove Press.

Fanon, Frantz. 2008. *Black Skin, White Masks*. Trans. Richard Philcox. New York: Grove Press.
Farr, James. 2008. "Locke, Natural Law, and New World Slavery." *Political Theory* 36, no. 4: 495–522.
Feagin, Joe R. 2010. *The White Racial Frame: Centuries of Racial Framing and Counter-Framing*. New York: Routledge.
Firmin, Anténor. 2002. *The Equality of the Human Races*. Trans. Asselin Charles. Urbana and Chicago: University of Illinois Press.
Flory, Dan. 2008. *Philosophy, Black Film,* Film Noir. University Park, PA: Pennsylvania State University Press.
Foucault, Michel. 2003. *"Society Must Be Defended": Lectures at the Collège de France, 1975–76*. Trans. David Macey. New York: Picador.
Fredrickson, George. 2002. *Racism: A Short History*. Princeton: Princeton University Press.
Frye, Marilyn. 2001 [1992]. "White Woman Feminist 1983-1992." In Boxill (2001b), pp. 83–100.
Garcia, J. L. A. 2012. [1996]. "The Heart of Racism." In Taylor (2012), vol. III, pp. 31–67.
Gines, Kathryn T. 2012. "Reflections on the Legacy and Future of the Continental Tradition with Regard to the Critical Philosophy of Race." *Southern Journal of Philosophy* 50, no. 2: 329–44.
Gines, Kathryn T. 2014. *Hannah Arendt and the Negro Question*. Bloomington: Indiana University Press.
Glasgow, Joshua. 2009a. "Racism as Disrespect." *Ethics* 120: 64–93.
Glasgow, Joshua. 2009b. *A Theory of Race*. New York: Routledge.
Gobineau, Arthur de. 1999 [1915]. *The Inequality of Human Races*. Trans. Adrian Collins. New York: Howard Fertig.
Goldberg, David Theo. 1993. *Racist Culture: Philosophy and the Politics of Meaning*. Cambridge, MA: Blackwell.
Goldberg, David Theo. 2002a. "Liberalism's Limits: Carlyle and Mill on 'The Negro Question'." In Ward and Lott, pp. 195–204.
Goldberg, David Theo. 2002b. *The Racial State*. Malden, MA: Blackwell.
Goldenberg, David M. 2003. *The Curse of Ham: Race and Slavery in Early Judaism, Christianity, and Islam*. Princeton, NJ: Princeton University Press.
Goldenberg, David M. 2009. "Racism, Color Symbolism, and Color Prejudice." In Eliav-Feldon et al., pp. 88–108.
Gooding-Williams, Robert. 2005. *Look, a Negro! Philosophical Essays on Race, Culture and Politics*. New York: Routledge.
Gooding-Williams, Robert. 2009. *In the Shadow of Du Bois: Afro-Modern Political Thought*. Cambridge, MA: Harvard University Press.
Gordon, Lewis R. 1995. *Bad Faith and Antiblack Racism*. Atlantic Highlands, NJ: Humanities Press.
Gordon, Lewis R., ed. 1997a. *Existence in Black: An Anthology of Black Existential Philosophy*. New York: Routledge.
Gordon, Lewis R. 1997b. *Her Majesty's Other Children: Sketches of Racism from a Neocolonial Age*. Lanham, MD: Rowman & Littlefield.
Gordon, Lewis R. 2000. *Existentia Africana: Understanding Africana Existential Thought*. New York: Routledge.
Gordon, Lewis R. 2008. *An Introduction to Africana Philosophy*. New York: Cambridge University Press.

Gossett, Thomas F. 1997. *Race: The History of an Idea in America*, new ed. Orig. ed. 1963. New York: Oxford University Press.
Gracia, Jorge J. E. 2005. *Surviving Race, Ethnicity, and Nationality: A Challenge for the Twenty-First Century*. Lanham, MD: Rowman & Littlefield.
Gracia, Jorge J. E., ed. 2007. *Race or Ethnicity? On Black and Latino Identity*. Ithaca, NY: Cornell University Press.
Graham, Kevin M. 2010. *Beyond Redistribution: White Supremacy and Racial Justice*. Lanham, MD: Rowman & Littlefield.
Hall, Barbara. 2005. "Race in Hobbes." In Valls (2005b), pp. 43–56.
Hannaford, Ivan. 1996. *Race: The History of an Idea in the West*. Baltimore: Johns Hopkins University Press.
Hardy, Paul-A. 2002. "Medieval Muslim Philosophers on Race." In Ward and Lott (2002), pp. 38–62.
Harris, Leonard, ed. 1999. *Racism*. Amherst, NY: Humanity Books.
Haslanger, Sally. 2000. "Feminism in Metaphysics: Negotiating the Natural." In Miranda Fricker and Jennifer Hornsby, eds. *The Cambridge Companion to Feminism in Philosophy*. New York: Cambridge University Press, pp. 107–26.
Haslanger, Sally. 2012. *Resisting Reality: Social Construction and Social Critique*. New York: Oxford University Press.
Hill, Jr., Thomas E., and Bernard Boxill. 2001. "Kant and Race." In Boxill (2001b), pp. 448–71.
Hobson, John M. 2012. *The Eurocentric Conception of World Politics: Western International Theory, 1760-2010*. New York: Cambridge University Press.
Hund, Wulf D. 2006. *Negative Vergesellschaftung: Dimensionen der Rassismusanalyse*. Münster: Verlag Westfälisches Dampfboot.
Hypatia: A Journal of Feminist Philosophy: 2007: Special Issue: Race and the Regulation of the Gendered Body 22, no. 2.
Isaac, Benjamin. 2004. *The Invention of Racism in Classical Antiquity*. Princeton, NJ: Princeton University Press.
Jeffers, Chike. 2013. "The Cultural Theory of Race: Yet Another Look at Du Bois's 'The Conservation of Races.'" *Ethics* 123: 403–26.
The Journal of Ethics: 2003: Special Issue: Race, Racism, and Reparations 7, no. 1.
The Journal of Social Philosophy: 2010: Special Issue: New Thinking in Race Theory 41, no. 3.
King, Richard H. 2004. *Race, Culture, and the Intellectuals, 1940-1970*. Baltimore, MD: Johns Hopkins University Press.
Kitcher, Philip. 2012 [2007]. "Does 'Race' Have a Future?" In Taylor (2012), vol. II, pp. 106–25.
Kleingeld, Pauline. 2007. "Kant's Second Thoughts on Race." *The Philosophical Quarterly* 57, no. 229: 573–92.
Lawson, Bill E., ed. 1992. *The Underclass Question*. Philadelphia: Temple University Press.
Lawson, Bill E., and Donald F. Koch, eds. 2004. *Pragmatism and the Problem of Race*. Bloomington: Indiana University Press.
Lee, Emily S., ed. 2014. *Living Alterities: Phenomenology, Embodiment, and Race*. Albany, NY: SUNY Press.
Levine, Michael P., and Tamas Pataki, eds. 2004. *Racism in Mind*. Ithaca, NY: Cornell University Press.
Lott, Tommy L. 1999. *The Invention of Race: Black Culture and the Politics of Representation*. Malden, MA: Blackwell.

Louden, Robert B. 2000. *Kant's Impure Ethics: From Rational Beings to Human Beings*. New York: Oxford University Press.
McCarthy, Thomas. 2009. *Race, Empire, and the Idea of Human Development*. New York: Cambridge University Press.
McGary, Howard, Jr. 1999. *Race and Social Justice*. Malden, MA: Blackwell.
McGary, Howard, Jr., and Bill Lawson. 1992. *Between Slavery and Freedom*. Bloomington: Indiana University Press.
McWhorter, Ladelle. 2009. *Racism and Sexual Oppression in Anglo-America: A Genealogy*. Bloomington: Indiana University Press.
Mallon, Ron. 2012 [2004]. "Passing, Traveling, and Reality: Social Construction and the Metaphysics of Race." In Taylor (2012), vol. II, pp. 53–81.
Mann, Anika Maaza. 2012 [2010]. "Race and Feminist Standpoint Theory." In Taylor (2012), vol. II, pp. 267–80.
Medina, José Medina. 2012. *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and Resistant Imaginations*. New York: Oxford University Press.
Mehta, Uday Singh. 1999. *Liberalism and Empire: A Study in Nineteenth-Century British Liberal Thought*. Chicago: University of Chicago Press.
Mills, Charles W. 1997. *The Racial Contract*. Ithaca, NY: Cornell University Press.
Mills, Charles W. 1998. *Blackness Visible: Essays on Philosophy and Race*. Ithaca, NY: Cornell University Press.
Mills, Charles W. 2005. "Kant's *Untermenschen*." In Valls (2005b), pp. 169–93.
The Monist: 2010: 93, no. 2: Race.
Moran III, Francis. 2002. "Between Primates and Primitives: Natural Man as the Missing Link in Rousseau's *Second Discourse*." In Ward and Lott, pp. 125–44.
Murphy, Joseph M. 1995. *Working the Spirit: Ceremonies of the African Diaspora*. Boston: Beacon Press.
Muthu, Sankar. 2003. *Enlightenment against Empire*. Princeton, NJ: Princeton University Press.
Ortega, Mariana, and Linda Martín Alcoff, eds. 2009. *Constructing the Nation: A Race and Nationalism Reader*. Albany, NY: SUNY Press.
Outlaw, Jr., Lucius T. 1996. *On Race and Philosophy*. New York: Routledge.
Outlaw, Jr., Lucius T. 2005. *Critical Social Theory in the Interests of Black Folk*. Lanham, MD: Rowman & Littlefield.
Pateman, Carole, and Charles W. Mills. 2007. *Contract and Domination*. Malden, MA: Polity.
Pitts, Jennifer. 2005. *A Turn to Empire: The Rise of Imperial Liberalism in Britain and France*. Princeton, NJ: Princeton University Press.
Roberts, Rodney, ed. 2002. *Injustice and Rectification*. New York; Peter Lang.
Sala-Molins, Louis. 2006. *Dark Side of the Light: Slavery and the French Enlightenment*. Trans. John Conteh-Morgan. Minneapolis: University of Minnesota Press.
Sartre, Jean-Paul. 2006. *Colonialism and Neocolonialism*. Trans. Azzedine Haddour, Steve Brewer, and Terry McWilliams. New York: Routledge.
Shelby, Tommie. 2004. "Race and Social Justice: Rawlsian Considerations." *Fordham Law Review* 72, no. 5: 1697–714.
Shelby, Tommie. 2005. *We Who Are Dark: The Philosophical Foundations of Black Solidarity*. Cambridge, MA: Harvard University Press.
Shelby, Tommie. 2012 [2002]. "Is Racism in the Heart?" In Taylor (2012), vol. III, pp. 68–77.
Sheth, Falguni A. 2009. *Toward a Political Philosophy of Race*. Albany, NY: SUNY Press.

South African Journal of Philosophy: 2011: "Symposium on Samantha Vice's Essay," 30, no. 4: 428–39.
Southern Journal of Philosophy: 2009: Supplement, Vol. 47: "Race, Racism, and Liberalism in the Twenty-First Century".
Strickland, Debra Higgs. 2003. *Saracens, Demons, and Jews: Making Monsters in Medieval Art*. Princeton, NJ: Princeton University Press.
Stubblefield, Anna. 2005. *Ethics along the Color Line*. Ithaca, NY: Cornell University Press.
Sullivan, Shannon. 2006. *Revealing Whiteness: The Unconscious Habits of Racial Privilege*. Bloomington: Indiana University Press.
Sullivan, Shannon, and Nancy Tuana, eds. 2007. *Race and Epistemologies of Ignorance*. Albany, NY: SUNY Press.
Sundquist, Eric J., ed. 1996. *The Oxford W. E. B. Du Bois Reader*. New York: Oxford University Press.
Sundstrom, Ronald R. 2008. *The Browning of America and the Evasion of Social Justice*. Albany, NY: SUNY Press.
Taiwo, Olufemi. 2010. *How Colonialism Preempted Modernity in Africa*. Bloomington: Indiana University Press.
Taylor, Paul, ed. 2012. *The Philosophy of Race*, 4 vols. New York: Routledge.
Taylor, Paul. 2013. *Race: A Philosophical Introduction*. 2nd ed. (Orig. ed. 2003). Malden, MA: Polity.
Tibebu, Teshale. 2011. *Hegel and the Third World: The Making of Eurocentrism in World History*. Syracuse, NY: Syracuse University Press.
Tully, James. 1994. "Rediscovering America: The *Two Treatises* and Aboriginal Rights." In G. A. J. Rogers, ed., *Locke's Philosophy: Content and Context*. London: Clarendon Press, pp. 165–96.
Valls, Andrew. 2005a. "'A Lousy Empirical Scientist': Reconsidering Hume's Racism." In Valls (2005b), pp. 127–49.
Valls, Andrew, ed. 2005b. *Race and Racism in Modern Philosophy*. Ithaca, NY: Cornell University Press.
Vasconcelos, José. 1997. *The Cosmic Race: A Bilingual Edition*. Trans. Didier T. Jaén. Baltimore, MD: Johns Hopkins University Press.
Vice, Samantha. 2012 [2010]. "How Do I Live in This Strange Place?" In Taylor (2012), vol. III, pp. 197–217.
Ward, Julie K., and Tommy L. Lott, eds. 2002. *Philosophers on Race: Critical Essays*. Malden, MA: Blackwell.
West, Cornel. 1982. *Prophesy Deliverance! An Afro-American Revolutionary Christianity*. Philadelphia: Westminster Press.
West, Cornel. 1989. *The American Evasion of Philosophy: A Genealogy of Pragmatism*. Madison: University of Wisconsin Press.
Yancy, George, ed. 2004. *What White Looks Like: African-American Philosophers on the Whiteness Question*. New York: Routledge.
Yancy, George. 2008. *Black Bodies, White Gazes: The Continuing Significance of Race*. Lanham, MD: Rowman & Littlefield.
Yancy, George, ed. 2010. *The Center Must Not Hold: White Women Philosophers on the Whiteness of Philosophy*. Lanham, MD: Rowman & Littlefield.
Yancy, George. 2012. *Look, a White! Philosophical Essays on Whiteness*. Philadelphia: Temple University Press.

Zack, Naomi. 1993. *Race and Mixed Race*. Philadelphia: Temple University Press.
Zack, Naomi, ed. 2000. *Women of Color and Philosophy*. Malden, MA: Blackwell.
Zack, Naomi. 2001 [1997]. "Race and Philosophic Meaning." In Boxill (2001b), pp. 43–57.
Zack, Naomi. 2011. *The Ethics and Mores of Race: Equality after the History of Philosophy*. Lanham, MD: Rowman & Littlefield.

Index of Names

Index