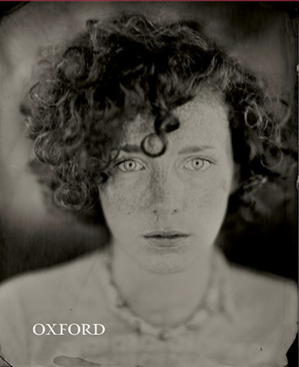
VOLUME 2

IMPLICIT BIAS & PHILOSOPHY

MORAL RESPONSIBILITY, STRUCTURAL INJUSTICE, AND ETHICS

EDITED BY
Michael Brownstein + Jennifer Saul



OXFORD

Implicit Bias and Philosophy

Volume 2: Moral Responsibility, Structural Injustice, and Ethics

Implicit Bias and Philosophy

VOLUME 2

Moral Responsibility, Structural Injustice, and Ethics

EDITED BY
Michael Brownstein
and Jennifer Saul

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

First Edition published in 2016
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: 20105947707

ISBN (Vol I) 978–0–19–871324–1
ISBN (Vol II) 978–0–19–876617–9
ISBN (Set) 978–0–19–876618–6

Printed in Great Britain by
Clays Ltd, St Ives plc

Contents

Part 1. Moral Responsibility for Implicit Bias

Part 2. Structural Injustice

Part 3. The Ethics of Implicit Bias: Theory and Practice

Contributors

Lawrence Blum, University of Massachusetts, Boston

Samantha Brennan, Western University

Michael Brownstein, John Jay College of Criminal Justice/CUNY

Luc Faucher, Université du Québec à Montréal

Joshua Glasgow, Sonoma State University

Katya Hosking, Pupil Barrister, Devereux Chambers

Anne Jacobson, University of Houston

Daniel Kelly, Purdue University

Clea F. Rees, Cardiff University

Roseanne Russell, Cardiff University

Jennifer Saul, University of Sheffield

Maureen Sie, Erasmus University Rotterdam and Leiden University

Nicole van Voorst Vader-Bours, Erasmus University Rotterdam

Natalia Washington, Washington University, St Louis

Robin Zheng, Yale-NUS College

Contents of Volume 1

Introduction

Michael Brownstein and Jennifer Saul

Several features of implicit biases demand attention from scholars working in moral and political theory, normative ethics, and applied ethics. One feature is the sheer pervasiveness of these biases. One review by Brian Nosek and colleagues (2007) of over 700,000 participants found that more than 70% of white participants demonstrate a moderate to strong preference for images of white faces over images of black faces on the race evaluation (or black-white) Implicit Association Test (IAT; Greenwald et al., 1998).[1] This finding is highly significant, given that millions of people have taken the race IAT so far, and that several reviews based on hundreds of studies have concluded that the race IAT is at least a moderate predictor of discriminatory behavior (Greenwald et al., 2003, 2009; Nosek et al., 2005, 2007; Lane et al., 2007). In addition to being pervasive, implicit biases are varied. While race and gender biases receive the most attention, the basic mechanisms of implicit social cognition have been shown to produce biased judgments of, and discriminatory behaviors toward, members of many social groups, such as disabled people, the elderly, and the LGBTQ community. While there is important heterogeneity within these kinds of implicit biases (see Chapters 1.3 and 1.4 in Volume 1), they are all similar in several respects that are relevant to ethical, moral, and political theorizing. First, each has the potential to be activated in particular contexts, and to affect agents' behavior, without the agent's conscious awareness. There is debate about exactly *how* unconscious implicit attitudes really are (see Chapters 1.3 and 1.4 in this volume), but there is no debate that implicit biases are different in at least some philosophically relevant respects from fully conscious, explicit biases. Second, implicit biases appear to be

[1] Roughly 40% of black participants prefer black faces over white faces on this IAT, 20% show no preference, and 40% prefer white faces over black faces (Nosek et al., 2002; Ashburn-Nardo et al., 2003; Dasgupta, 2004).

impressively resistant to agents' "direct" efforts to change or control them via force of will or mere intention. There *is* evidence, however, that implicit biases are malleable, and this evidence has been put into service in developing a number of techniques for the self-regulation of implicit bias. These empirical discoveries give rise to a range of important philosophical questions; in particular, whether, and to what extent people are responsible for their biases; the relationship between implicit biases and structural injustices; and, of course, what sort of ethical schema(s) are most effective for combating bias.

1 Part 1: Moral Responsibility for Implicit Bias

The first of the three parts of this volume is dedicated to the topic of moral responsibility for implicit bias. Natalia Washington and Daniel Kelly argue that moral responsibility for implicit bias turns not on whether the holder of the bias knows that they are biased, but rather on whether this knowledge is available in their community. Joshua Glasgow argues, somewhat surprisingly, that responsibility for an implicit bias may depend in part on the *content* of the attitude. Robin Zheng suggests that we obtain a more illuminating framework for considering implicit bias if we distinguish between "attributability" and "accountability." Maureen Sie and Nicole van Voorst Vader-Bours urge a switch to a focus on collective rather than individual responsibility. And finally, Luc Faucher argues that reflection on implicit biases gives us good reason to revise (though not abandon) our standard philosophical conceptions of moral responsibility. Together, these papers offer us a rich new range of options for understanding moral responsibility for implicit biases, and indeed, moral responsibility as such.

In Chapter 1.1, Washington and Kelly argue that there are cases in which a person should be held responsible for behaviors influenced by implicit bias, even if the person is unaware of being biased. They frame the issue in terms of control-based and knowledge-based exculpating conditions that are commonly taken to excuse bad but not blameworthy actions. Focusing on knowledge-based exculpating conditions in particular, Washington and Kelly argue that the kind of knowledge relevant to moral responsibility and exculpation need not be "in the head" of the agent whose actions is being evaluated. Rather, the kind of knowledge relevant to moral responsibility and exculpation is knowledge available in one's epistemic environment. This indexes moral responsibility to what one *should* know, given one's epistemic environment. Washington and Kelly develop this claim in terms of common forms of moral reasoning in analogous cases, as well as in terms of "externalist" conceptions of knowledge.

Glasgow takes as his starting point the fact that people feel alienated from their implicit biases, treating this as a key datum in need of explanation. This datum, Glasgow points out, is in tension with the fact that people feel guilty about their implicit biases, even if they genuinely and wholeheartedly disavow them. The feeling of guilt involves not just embarrassment, on Glasgow's view, but also the sense that one is *worthy* of embarrassment. But this judgment of worthiness is puzzling if we truly disavow biased attitudes and judgments. After critically examining a number of alternative accounts, Glasgow settles on the view that responsibility attributions can shift depending on the content of the action or attitude, and that the criterion that determines the shift is the harm done in or by the act or attitude in question.

In Chapter 1.3, Zheng focuses on the distinction between responsibility as "attributability" and responsibility as "accountability." While attributability has to do with whether an attitude or action is an expression of one's agency, accountability refers to the social and institutional practices that govern when it is appropriate for people to bear the consequences of their actions. Corresponding to these two conceptions of moral responsibility are two types of interpersonal responses, Zheng argues: appraisal-based responses (e.g. blame) and non-appraisal-based responses (e.g. penalties). On Zheng's view, a person can lack attributability for an implicit bias, but still be held accountable for the effects of that bias on others. For both pragmatic and moral reasons, Zheng argues, we ought to refrain from offering judgmental appraisal-based responses to others in response to their implicitly biased behaviors. Instead, we ought to focus on accountability and non-appraisal-based responses, which seek to address the problem at hand without making pronouncements on the character of the people involved. Carefully distinguishing between attributability and accountability not only does justice to our moral experience and agency, Zheng argues, but is also more effective in reducing the prevalence and impact of implicit biases.

A different approach is charted in Chapter 1.4 by Sie and Voorst Vader-Bours. They advance an "indirect" argument for the plausibility of moral responsibility for behaviors affected by disavowed stereotypes and prejudices. Their claim is indirect in the sense that it is focused on collective rather than individual action—in particular, the ability of individuals to contribute to collective efforts to combat discrimination. Sie and Vader-Bours also distinguish between taking responsibility for some action and "self-ascribing" responsibility. The latter involves critical acts of self-scrutiny, which they describe through a series of examples.

Finally, in Chapter 1.5—the last chapter of the first part of this volume—Faucher takes a broader look at the plausible ways in which the data on implicit bias might cause philosophers to revise their conceptions of moral responsibility. The conclusion of this chapter is that while the existence of implicit bias forces philosophers to rethink key components of traditional conceptions of moral responsibility, it is not cause for the abandonment of the concept of moral responsibility altogether. Faucher provides a taxonomy of forms of revisionism and argues for what Manuel Vargas (2005) has called "moderate" and "sophisticated" revisionism. This entails dropping some, but not all, elements of our standard philosophical ways of thinking about moral responsibility. In particular, Faucher argues that the data on implicit bias require that we rethink the control and awareness conditions for moral responsibility, and that we also reconsider the notion that agents can only be responsible for actions that express their "real" or "deep" self.

2 Part 2: Structural Injustice

The second part of this volume moves beyond questions about moral responsibility to consider the connections between the attitudes held by individuals and the structural injustice of the societies in which they are situated. Lawrence Blum criticizes a widespread focus on individual psychology, arguing that an individualistic approach to stereotype threat can help to mask structural injustices and to depoliticize what should be seen as political matters. And Anne Jacobson compares individual and collective responses to implicit bias through a series of wideranging analogies, also arguing for the importance of a non-individualistic approach.

In Chapter 2.1, Blum elaborates on the importance of considering structural injustice within the study of stereotype threat, a phenomenon in which, under certain circumstances, individuals' performance at particular tasks may suffer due to awareness of widely held stereotypes about their group. Blum argues that Claude Steele's framework for understanding stereotype threat fails to distinguish clearly between sound generalizations and stereotypes as evidence-resistant overgeneralizations. This, Blum argues, may discourage the forming of the accurate generalizations that are essential in diagnosing disparities between groups (e.g. in educational performance). In doing so, Blum claims, Steele's framework masks the asymmetries in vulnerability to stereotyping that generate structural injustices. The masking of these asymmetries is connected with Steele's view that vulnerable groups, such as black students, may not have internalized the stereotypes about their group. It is also connected, Blum concludes, to the lack of political and civic perspective in the analysis of stereotype threat and to common depoliticized suggestions for reducing it.

In Chapter 2.2, Jacobson articulates concerns about the individualism found in research on implicit bias. Reviewing a wide range of research, from work in neuroscience to health care, she argues that neither attitude change nor institutional change alone can effectively combat intergroup discrimination. Jacobson emphasizes in particular the many complexities and potentially unforeseen consequences of any social change effort.

3 Part 3: The Ethics of Implicit Bias: Theory and Practice

The final part of this volume takes a broader perspective on the ethics of implicit bias. Clea Rees begins by arguing that we need to focus on how to become more virtuous agents who are able to control our implicit biases—a task which requires an appropriate moral education and community. Michael Brownstein hones in on the underexplored contextual variation in the manifestation of implicit bias, exploring its implications for character and also for combating implicit biases. Samantha Brennan tackles the fact that the harms caused by implicit biases will often (though not always) be very small ones—a kind of harm that moral theorists are notably poor at thinking about. Finally, Katya Hosking and Rosalind Russell turn to the very practical issue of how discrimination law should treat implicit biases, taking UK law as their example.

In Chapter 3.1, Rees brings together several themes from both this and the previous volume. She argues that effectively combating implicit bias requires strong individual commitments to treat people fairly, as opposed to merely holding strong egalitarian beliefs. Rees draws on research showing that when such strong commitments are automatised, they affect cognitive-affective processing without conscious effort or awareness. Such commitments are not only free from many of the limitations which constrain the potential effectiveness of other self-regulation strategies, but they may also gradually alter the underlying cognitive-affective processing system, thus weakening implicit biases themselves. Rees concludes by arguing that the development and maintenance of individual commitments to fairness is not a purely individual matter. As Aristotle said, individual virtue requires an education and community which actively encourages virtuous habituation as well as thoughtful deliberation.

In Chapter 3.2, Brownstein considers the ethical ramifications of the effects of context on the activation and expression in behaviour of implicit biases. He argues that an ethics of implicit bias must focus on outlining how agents can cultivate the right sort of relationships with the situations and contexts that affect their attitudes and behavior. This notion, of cultivating the right sort of "ambient" relationships, has been underdescribed by most ethical thinking about implicit bias, he argues. Such discussions usually focus on the relationship between attitudes or mental states *within* agents, not considering their interaction with context. Brownstein discusses strategies found in the attitude change literature that illustrate a "contextualist" ethics. These contextualist strategies, he argues, help to point the way forward in designing more effective bias-reduction interventions.

In Chapter 3.3, Brennan examines the moral significance of micro-inequities. She explains why the phenomena of micro-inequities and implicit bias often go together, and carefully explores the ways in which it is and is not appropriate to connect them. Brennan then examines the various sorts of analyses available for understanding the moral significance of micro-inequities, and argues against skeptical responses to the claim that they are morally significant.

The final chapter of the book considers an applied problem: how does anti-discrimination law treat implicit biases, and how ought it to treat them? Russell and Hosking examine discrimination law in the UK, arguing that the current provisions in the Equality Act 2010 are far more limited in practice than the statutory language suggests. The problem is that, despite the strong indications of disadvantage and exclusion due to implicit bias at a statistical level, it is virtually never possible to say for certain that implicit bias was a cause of a particular individual outcome. This is a problem with fault-based models of discrimination, which allow implicit bias to operate without a remedy. Russell and Hosking then argue that the proactive model of equality law developed in the public sector can offer distinct advantages. In contrast to discrimination law, the public sector equality duty sidesteps questions of individual fault by imposing a positive obligation on public sector bodies to "have due regard" to the need to eliminate discrimination and advance equality of opportunity. These bodies are also required to analyse and publish information about their performance in relation to this duty, to engage with members of protected groups, and to publish equality objectives which they have set for themselves on the basis of this evidence. Sadly, however, the promise of equality law largely fails to deliver in practice. This is due, Russell and Hosking argue, to a broad "folk-liberal" orientation in UK law and legislative processes. This orientation downplays the structural character of discrimination while elevating the formal value of liberty. Arguments for formal equality of opportunity fit comfortably within this framework, but arguments for substantive equality do not; yet it is a commitment to substantive equality which is needed to begin tackling the effects of implicit bias.

All of the chapters in this volume are of clear relevance to philosophers working in moral and political theory, normative ethics, and applied ethics. As in the previous volume, each chapter integrates critical philosophical thinking with the latest empirical data, and each chapter meets this challenge in a unique way. While some of the chapters operate closely within the bounds of familiar philosophical concepts, testing how the empirical data meet relevant philosophical conditions, other chapters propose substantive revisions to familiar philosophical concepts on the basis of emerging data. A particularly important feature of all of these chapters is that, though clearly philosophical, they never lose sight of the ultimate goal of creating a more just world.

Acknowledgements

Michael Brownstein and Jennifer Saul are extremely grateful for each contribution to these volumes and to the valuable feedback provided by anonymous reviewers for Oxford University Press. Many thanks also to the OUP team: Peter Momtchiloff, Eleanor Collins, Manikandan Chandrasekaran, R.A. Marriot, and Hayley Buckley. We are also grateful to Keliy-Anderson Staley for providing the cover images, to Raymond Drainville for designing the covers, and to Lacey Davidson and Isabel Gois for creating the indexes. The Implicit Bias and Philosophy workshops at the University of Sheffield, from which these volumes emerged, would not have been possible without the generous support of the Leverhulme Trust and the expert organization of Angela Pepper. We would like to thank all the participants at these conferences for the fantastic discussions. Finally, we cannot thank our families enough. All our thanks to Reine Hewitt, Leda, Iggy, and Minerva Brownstein, and Raymond Drainville and Theo Drainville-Saul.

References

Ashburn-Nardo, L., Knowles, M., and Monteith, M. (2003). “Black Americans’ implicit racial associations and their implications for intergroup judgment.” *Social Cognition* 21(1): 61–87.

Dasgupta, N. (2004). “Implicit ingroup davoritism, outgroup favoritism, and their behavioral manifestations.” *Social Justice Research* 17(2): 143–68.

Greenwald, A. G., Nosek, B., and Banaji, M. (2003). “Understanding and using the Implicit Association Test: I. An improved scoring algorithm,” *Journal of Personality and Social Psychology* 85(2): 197–216.

Greenwald, A. G., Poehlman, T., Uhlmann, E., and Banaji, M. (2009). “Understanding and Using the Implicit Association Test: III Meta-Analysis of Predictive Validity.” *Journal of Personality and Social Psychology* 97(1): 17–41.

Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. K. (1998). “Measuring individual differences in implicit cognition: The Implicit Association Test.” *Journal of Personality and Social Psychology* 74: 1464–80.

Lane, K., Kang, J., and Banaji, M. (2007). "Implicit social cognition and law." *Annual Review of Law and Social Science* 3: 427–51.

Nosek, B., Banaji, M., and Greenwald, A. G. (2002). "Harvesting intergroup implicit attitudes and beliefs from a demonstration website." *Group Dynamics* 6: 101–15.

Nosek, B., Greenwald, A. G., and Banaji, M. (2005). "Understanding and using the Implicit Association Test: II. Method variables and construct validity." *Personality and Social Psychology Bulletin* 31(2): 166–80.

Nosek, B., Greenwald, A. G., and Banaji, M. (2007). "The Implicit Association Test at age 7: A methodological and conceptual review." In Bargh, J. A. (ed.), *Automatic Processes in Social Thinking and Behavior*. Philadelphia, PA: Psychology Press.

Vargas, M. (2005). "The revisionist's guide to responsibility." *Philosophical Studies* 125(3): 399–429.

PART 1

Moral Responsibility for Implicit Bias

1.1

Who's Responsible for This?

Moral Responsibility, Externalism, and Knowledge about Implicit Bias

Natalia Washington and Daniel Kelly

1 The Cognitive Monster

Recently, philosophers have become increasingly concerned about a cluster of issues that arise at the intersection of ethics and psychology. The general worry was expressed by John Bargh in his influential paper "The Cognitive Monster" (1999):

> If it were indeed the case, as research appeared to indicate, that stereotyping occurs without an individual's awareness or intention, then the implications for society—specifically, the hope that prejudice and discrimination could eventually be eradicated—were tremendous, as well as tremendously depressing. Most ominously, how could anyone be held responsible, legally or otherwise, for discriminatory or prejudicial behavior when psychological science had shown such effects to occur unintentionally? (363)

We agree with Bargh that the picture emerging from many areas of empirical psychology is both theoretically intriguing and morally troubling. Taken as a whole, he sees this picture as suggesting that a significant amount of human behavior is at the behest of a cognitive monster that operates outside of our conscious awareness, producing behaviors unguided by explicit intention. When they are produced in such a way, it is difficult to see how we could justifiably be held responsible for those behaviors, given the principles that typically govern practices of holding people responsible. On the one hand, facts about the kind of implicit and automatic mental processes that Bargh alludes to in his description of the "cognitive monster" do not fit easily with a folk psychological picture of the mind, and as we will see, the way that they deviate from that picture can make them seem not just counterintuitive but somewhat unsettling. On the other hand, theories of moral responsibility are often grounded in intuitions, social norms, and practices that rely on commonsense conceptions of the sources of behavior. This raises a difficulty concerning how to square these facts with these theories, and how the latter can best be brought to bear on the former. In this paper we aim to think systematically about, formulate, and begin addressing some of the challenges to applying theories of moral responsibility to behaviors shaped by a particular subset of unsettling psychological complexities: namely, implicit biases.[1]

One might initially be skeptical that implicit biases raise any special challenge to moral responsibility. We disagree. Echoing some of Bargh's themes, Jennifer Saul (2013) sketches a position about implicit bias and blameworthiness in terms of awareness and control:

> I think it is also important to abandon the view that all biases against stigmatized groups are blameworthy. My first reason for abandoning this view is its falsehood. A person should not be blamed for an implicit bias that they are completely unaware of, which results solely from the fact that they live in a sexist culture. Even once they become aware that they are likely to have implicit biases, they do not instantly become able to control their biases, and so they should not be blamed for them. (55)

Saul's discussion shows that there are intricacies here that deserve careful philosophical attention. Indeed, we follow her in framing our questions about responsibility and blame in terms of control and awareness or knowledge.[2] Moreover, we agree with her more specific claim that individuals do not become able to control their biases simply by or immediately upon learning about them. The case we will make below takes this as given, and argues that there are nevertheless cases in which an individual should be held responsible for actions that are influenced by her implicit biases—even if she cannot control them at the time of the behavior, and even if she does not know she has those implicit biases, and would disavow those biases were she under their influence.[3] For reasons that will

[1] Implicit bias hardly exhausts the domain of interesting, counterintuitive, and philosophically relevant "unsettling facts" being discovered about human psychology. In order to focus our discussion in this paper we will be bracketing some other relevant areas of research, including those on the unreliability of self-report (Nisbett and Wilson, 1977; Wilson, 2002), confabulation and post hoc rationalization (Hirstein, 2005; Haidt, 2006; Tavris and Aronson, 2007), and stereotypes and stereotype threat (Carr and Steele, 2010).

[2] See Kelly and Roedder (2008: 532) for similar worries expressed in terms of Frankfurtian identification and Fischer and Ravizza's notion of reasons-responsiveness.

[3] In a recent paper, Jules Holroyd (2012) contends that arguments in support of the conclusion that individuals cannot be held responsible for manifesting implicit biases are untenable. We want to make the stronger claim, that in many situations, individuals *are* responsible for manifesting implicit biases. We owe much to earlier discussions with her on this subject.

become evident as we go, our discussion will focus not on control but rather on knowledge about implicit biases. We should also be clear that the view we articulate and defend applies first and foremost to actions, rather than the agent who engages in an action. For the purposes of this essay we will remain neutral on the issue of whether or not someone is a bad person merely in virtue of harboring implicit biases, or if simply having implicit biases should be reflected in assessments of that person's character. Rather, we are concerned with whether and how a person should be held responsible when her biases are allowed to manifest in a specific piece of behavior. For us, the primary target of evaluation is action itself, even though responsibility and blame are eventually ascribed to the agent who engages in it.

Also central to our framing of questions about responsibility will be the notion of an exculpating condition. Broadly speaking, exculpating conditions specify factors that can excuse a person for an action, absolving her of responsibility and blame. In Section 2 we explain how such conditions operate in the common norms and practices that govern how we typically hold people responsible for some behaviors, but let them off the hook for others. Next we lay out several core features of implicit bias, emphasizing those features that correspond to common exculpating conditions, and spelling out how such correspondences motivate the kinds of worries about responsibility expressed by Bargh and explored by Saul. We then present a thought experiment designed to show that in certain circumstances, an individual who does not know she has implicit biases can nevertheless be held responsible for behaviors that were crucially influenced by those implicit biases. In reflecting on the intuition our thought experiment is designed to pump[4], we note that it implies that not all of the knowledge relevant to moral responsibility and exculpation need be "in the head" of the individual whose actions are being evaluated. Hence, one of our main claims is that an individual can be open to blame for manifesting implicit biases when knowledge about such mental states is available in her epistemic environment, and that individual occupies a social role to which implicit biases and knowledge about them are clearly relevant. Since we see ourselves as articulating the best way to extend current practices concerning responsibility to actions that involve the psychological complexities of the cognitive monster, along the way we illustrate and defend this claim by comparing our thought experiment to a number of parallel but more familiar cases. In Section 6 we respond to some common objections and comment on the context and pragmatic implications of our position.

[4] See Dennett (2013) for a discussion of thought experiments as "intuition pumps."

2 Holding Ourselves Morally Responsible: Common Practices, Typical Excuses

In broad strokes, it is common practice to hold people morally responsible for many of their actions, where being held responsible is tied up with notions of praise and blame, reward, and punishment. When a person performs an action that is morally good, particularly when it is exemplary or supererogatory, the person may be praised, and can be justifiably rewarded. More importantly, when a person performs an action that is morally bad or wrong, the person is open to blame, and may be justifiably punished.[5]

Of course, the picture is not so simple. For instance, people are not held morally responsible for all of their behaviors. It is also common practice to excuse people for behaviors in special kinds of circumstances. Broadly speaking, such circumstances include those in which the behavior in question is forced or coerced, in which the behavior is accidental, or behaviors in which the agent is ignorant or unaware of some key element of her situation.[6] In these cases, we typically do not praise or reward the agent, even if the behavior has a morally good or desirable outcome, nor do we blame or punish the agent if the behavior has a morally bad or undesirable outcome. Behaviors that do not occur in such circumstances are sometimes said to be free. They can be described by phrases like "the behavior was freely chosen" or "it was the result of a genuine decision;" in some relevant sense, the agent could have reasonably done otherwise. In short, we hold each other morally responsible for those actions that are freely chosen—expressions of free will.[7]

[5] It is common practice in the cultural environment that the authors and probably most readers of this chapter inhabit. Whether even this very general characterization of the norms and practices concerning ascription of responsibility and blame applies to all cultures, or merely to those cultures that are WEIRD or heavily influenced by WEIRD cultures, is a fascinating and still largely underexplored question (Henrich et al., 2010; though see Sommers, 2011).

[6] Although we mention here only control- and knowledge-based exculpation conditions, we do not assume these are the only two types; for instance, see Machery et al. (2010) for discussion of a rationality-based condition.

[7] We realize we are passing over some rich philosophical ground quickly in these paragraphs, but we will complicate the picture as needed to make our key points as we go. Here we simply wish to give a coarse-grained overview of the types of connections between the key concepts of free will, responsibility, and praise and blame, as they are construed by folk psychology and common practice (or at least as how many "Introduction to Philosophy" courses depict folk psychology and common practices as construing those connections.) For discussion of some more fine-grained notions related to responsibility that philosophers have advanced, see Watson (1996), Shoemaker (2011), and Smith (2012). Most of what we talk about in terms of responsibility seems to us to most comfortably fall under the category of "accountability." However, little appears to be settled in this area (cf. Sripada, forthcoming). Even if the distinct notions prove to be defensible and significant, questions about if

Philosophers interested in these issues have long been concerned about the possibility of global threats to free will and moral responsibility. For instance, one such threat seems to emerge from contemporary physics, and an appreciation of the fact that there seems to be no room for genuine choice or moral responsibility if we are living in a deterministic universe. Others worry that another sweeping threat may be looming in the results of recent cognitive neuroscience which suggests that our actions are the result of mental processes that completely bypass our conscious, reflective deliberation and decision making (see Nahmias 2010, Roskies 2006). Although fascinating, our concern here will not be with either of these "global challenges" to moral responsibility. Rather, we will adopt the assumption built into our ordinary, everyday practices of holding people responsible—typical of compatibilist approaches to free will—that we *are* responsible for *some* of our behaviors.

To illustrate, consider this piece of behavior: Cate eats a batch of the cookies that her roommate made especially for tomorrow's bake sale. In the first variation of this scenario, Cate knows full well that her roommate made the cookies especially for the bake sale, but eats them anyway, for no other reason than because she is hungry. Described as such, Cate's behavior is not just callous, but it is of the sort for which she is responsible; she is a straightforward target for blame. However, in other circumstances we would not hold her responsible for the same piece of behavior. In a second variation of this scenario, imagine that Cate eats a batch of the cookies that her roommate made especially for tomorrow's bake sale, but she does so while in a somnambulant daze. Since in this case there is an important sense in which she did not know what she was doing, it is not her fault for ruining her roommate's contribution to the bake sale, and she should not be blamed but excused, because people are not commonly held responsible for their behaviors while they are sleep-walking.

Generalizing from this second variation, we will say that behaviors that occur in these kinds of scenario satisfy an *exculpating condition*. We will focus on two of the most important kinds of exculpating condition: those that center on knowledge and control, respectively. To a first approximation, in cases that satisfy the knowledge condition, the agent is excused for the behavior because she did not know or was unaware of relevant features of the situation, and in cases that satisfy the control condition, the agent is excused for the behavior because she did not have the right kind of control over it. This makes sleep-eating Cate an especially apt case for exculpation, since at first blush it seems that she is both unaware that

and how each one applies to the kinds of cases we consider here, while potentially interesting and important, would require more space than we have here to address properly.

she is sleep-eating, and unable to stop herself. In any event, we can begin articulating this line of thought as follows:

Knowledge Condition: An agent is exculpated for having done X if
(1) she did not know that she did X, or
(2) she did not know why she did X.

Control Condition: An agent is exculpated for having done X if
(1) she was constrained or coerced to do X, or
(2) she lacks proper control over doing X.

In addition to letting people off the hook for what they do in certain circumstances, ordinary practices of ascribing responsibility also allow for another wrinkle; in our terminology, exculpating conditions also have exception clauses. When a case of behavior occurs in circumstances that satisfy an exculpating condition, but the circumstances *also* meet one of the exculpating condition's exception clauses, then the agent is not excused, but instead is held responsible for the action. Consider a third variation on our cookie thief: Cate, while in a somnambulant daze, eats a batch of the cookies that her roommate made especially for tomorrow's bake sale, but she was in that somnambulant daze because she had taken a hefty dose of Ambien. Cate has a long history, of which she and her roommate are both well aware, of sleep-walking and binge eating whenever she takes Ambien. In this case, Cate's behavior satisfies an exculpating condition (indeed, a case could be made that she satisfies both), but it also satisfies an exception clause (or two): somnambulant Cate is unaware of and unable to consciously control what she is doing, but, like a drunk driver, she is responsible for having put herself in that compromised condition, and is blameworthy for what she does once she inhabits it—in this case, especially since she has a well-known history of such Ambien-induced destructive behavior.[8] Excuses, as they say, wear thin. More generally, exception clauses can be represented like this:

Knowledge Condition: An agent is exculpated for having done X if
(1) she did not know that she did X (except when she is responsible for having been unaware of doing X), or
(2) she did not know why she did X (except when she is responsible for having been ignorant)

[8] For more on these kinds of so-called "tracing" cases, see Vargas (2005) and Fischer and Tognazzini (2009), and for a discussion focused on culpable ignorance, see Smith (2011).

Control Condition: An agent is exculpated for having done X if
(1) she was constrained or coerced to do X (except when she is responsible for having been constrained), or
(2) she did not have volition or control over doing X (except when she is responsible for lacking that control)

We will return to this framework in Section 4, where we will show how it applies to and illuminates a few more fairly mundane cases, before bringing it to bear on a case that involves implicit bias. First, however, we will briefly describe how we are construing this range of unsettling psychological facts.

3 "Textbook" Facts about Implicit Bias

Much is still being discovered about the character of implicit biases, and we do not wish our argument to depend on any of the more controversial or uncertain aspects of the ongoing research. To that end, we will work with a fairly broad conception of implicit biases as unconscious and automatic negative evaluative tendencies directed towards people based on their membership in a stigmatized social group—for example, on gender, sexual orientation, race, age, or weight. Such biases appear to be widespread in many populations, cultures, and countries. For ease of exposition and to focus the discussion to come, we will confine most of our attention to implicit racial biases. We hope and suspect that what we have to say generalizes straightforwardly to other types of implicit biases as well.[9]

Here are four features of implicit bias that will be important for the discussion to come:

1) *Dissociation* In a single individual, implicit racial biases can coexist with explicit racial attitudes that are diametrically opposed to them. For example, a person can explicitly hold genuine anti-racist or egalitarian views that they sincerely endorse upon reflection, and yet at the same time harbor implicit biases again members of certain races.
2) *Introspective opacity* Typically, a person who is explicitly biased knowingly and intentionally evaluates others negatively based on their race. In contrast, a person who is only implicitly biased has tendencies whose presence and influence on thought and behavior is not easily detectable via introspection.

[9] For citations and details, see Brownstein and Saul's Introduction in Volume 1. The view of implicit biases articulated there is fairly standard, and bears much in common with the minimal view on which we rely. For more detailed discussion see the chapters by Frankish, Huebner, Holroyd and Sweetman, Machery, and Mallon in the first part of that volume, as well as Staats and Patton (2013) and Banaji and Greenwald (2013).

Moreover, her sincere self-reports about her own attitudes are not likely to reflect her implicit tendencies. Indeed, much of what is known about implicit biases comes not from self-report, but is rather inferred from indirect experimental techniques like the Implicit Association Test, startle eye blink tests, and semantic priming tests. These experimental techniques are indirect in that they do not directly rely on participants' powers of introspection or the accuracy of descriptions of their own psychological makeup.

3) *Recalcitrance* Implicit biases operate not just implicitly but automatically. It is difficult to completely suppress the manifestation of an implicit bias in either judgment or behavior. They are also much easier to acquire than they are to eradicate or completely remove from one's psychological makeup. Moreover, while they are amenable to some methods of control, directly suppressing their expression once activated requires vigilance and effort, is mentally fatiguing, and can backfire in a number of ironic ways.[10]
4) *Widespread effects on behavior* Implicit racial biases can influence judgments and behaviors in subtle but important ways, even in real world situations. Many studies suggest implicit biases influence snap decisions, such as determining whether or not a person is holding a weapon (Payne, 2005, 2006), or if a basketball player has committed a foul (Price and Wolfers, 2010). There is reason to believe that implicit biases can also influence more deliberate, temporally extended decision making, despite confidence that in such cases behavior is more likely to reflect explicit attitudes and considered views. Examples include what diagnosis or type of health care a medical patient should get (Blair et al., 2011), who should or should not to serve on a jury (Haney-López, 2000), and whom to hire, or which resumé gets an interview (Bertrand and Mullainathan, 2004; see also Kawakami et al., 2007).

Given these four features, we can further refine the worry behind Bargh's invocation of a cognitive monster. We have formulated two core exculpating conditions: one for knowledge and one for control. These conditions align neatly with two prominent properties of implicit biases: namely, that their existence and influence is not easily detectable via introspection—an individual can have and be

[10] Happily, it seems that implicit biases are not completely intractable or uncontrollable. Work on what we will call the malleability of implicit biases has shown a number of techniques to be quite promising, including both implementation intentions and exposure to counterstereotypical exemplars. For further discussion on the implications of this work, see Madva (ms.) and the chapters by Rees and Brownstein in these volumes.

influenced by implicit biases without *knowing* it—and that they are recalcitrant, liable to run automatically in the face of contradictory, explicitly held beliefs—an individual's implicit biases operate without and sometimes beyond her *control*. Given this, the potential trouble can be put rather starkly: behaviors driven by implicit biases will, in virtue of that neat alignment, satisfy one or both of the two key exculpating conditions, so the agent who engages in implicit bias-driven behavior should be absolved of responsibility or blame for it. Expressed another way:

Epistemic Worry Since implicit biases are opaque to introspection and can operate outside of conscious awareness, a person should be exculpated, and not blamed or held responsible for behaviors that manifest them.

Bypassing Worry Since the operation of implicit biases is recalcitrant and automatic, a person should be exculpated, and not blamed or held responsible for behaviors that manifest them.

In Sections 4 and 5 we spell out one way to dispel these two worries generated by the neat alignment between these exculpating conditions and this set of unsettling psychological facts. We argue that the growing knowledge about implicit biases, which includes a body of empirical research showing how implicit biases can best be brought under control, has important implications for responsibility and blame. Indeed, we suggest how this knowledge should be incorporated into what Manuel Vargas (2013) calls our *moral ecology*. We do this by showing how that research and its dissemination are relevant to the kinds of exception clauses and exculpating conditions discussed above.

4 The Hiring Committee

Consider three different people, each with one of the following psychological profiles:[11]

The Earnest Explicit Racist (circa whenever) The earnest explicit racist has implicit racial biases, but these are accompanied by explicitly racist attitudes as well. She is fully aware that she holds these explicit views and is able and willing to articulate them, though perhaps only among trusted friends. She reflectively endorses her racist attitudes, and acts on them without compunction when given the chance. Though she does not know about her implicit biases, if made aware she would take pride in the fact that these instinctive evaluative tendencies run in tandem with her more reflective judgments, and that both express her considered values.

[11] We are not under the mistaken impression that these characters exhaust the range of possible or interesting cases when it comes to evaluating implicitly biased behavior. We have chosen these three to highlight a particular situational feature, and to make the case that people can be blamed for behaviors driven by implicit biases they do not know they have. In Section 6 we will briefly consider some questions about how to extend our approach to other cases.

The Old-School Egalitarian (circa 1980) The old-school egalitarian is explicitly anti-racist. She genuinely holds egalitarian views, which she honestly reports when asked about her views on race, and which she sincerely endorses upon reflection. However, the old-school egalitarian also harbors implicit racial biases. Like almost everyone in 1980, though, she does not know this fact about herself. Not only is she unaware that she herself is implicitly biased, she has never heard of implicit biases at all. Unlike the explicit racist, she suffers from what we called dissociation, and so if it were somehow revealed to her that she was implicitly biased, she would be surprised and taken aback. In fact, she would disavow those evaluative tendencies, and would acknowledge that in cases where her implicit racial biases influenced her decisions or behaviors, something had gone wrong. Such decisions and behaviors would not express her considered values, and she would be falling short of her own avowed ideals.

The New Egalitarian (circa 2014) Like the old-school egalitarian, the new egalitarian is genuinely explicitly anti-racist and egalitarian. He too harbors implicit racial biases but does not know this fact about himself. Like many others in 2014, however, the new egalitarian is vaguely aware of the phenomenon of implicit bias, but has not looked into the matter very much, and so does not know any details. Nor has he checked to see if he has any implicit biases himself. As a result, he takes no precautions and makes no adjustments to his own behavior to suppress or counteract them. However, as with the old-school egalitarian, if the new egalitarian came to know that he was implicitly biased he would not endorse those evaluative tendencies, but would sincerely disavow them. He too would acknowledge that in cases where his implicit racial biases influenced his decisions or behaviors, something had gone wrong. He would be failing to express his values and to live up to his own avowed ideals.

Now indulge us in a little bit of science fiction fancifulness (given how the members of our cast of characters are indexed to different times), and imagine a scenario where these three people comprise a hiring committee. They have all the usual duties that come with membership on such committees, but most important is their task of sorting through the set of resumés submitted for consideration for the job, reading them over with an eye toward deciding which candidates to interview and, ultimately, to hire. As such, each committee member puts in the considerable time it takes to sort and evaluate those resumés, and the individual and collective decision-making processes both involve lots of conscious, explicit, and deliberate reasoning. But the processes are all unknowingly influenced by the implicit racial biases of the individual members as well. As a result, all of those selected to be interviewed turn out to be white; the committee overwhelmingly favored resumés from candidates with “white-sounding” names, even though there were equally well-qualified candidates of many racial and ethnic backgrounds in the application pool (the kind of outcome found in e.g. Bertrand and Mullainathan, 2004).

Now we are able to pose our paper’s eponymous question: Who is responsible for this? The outcome is clearly unjust, as many deserving candidates were not given a fair shot at the job for reasons that had nothing to do with their qualifications. Where does blame attach, and how blameworthy is each member of the committee for bringing about this outcome?

It seems to us that the earnest explicit racist is the most straightforward case, and thus the least interesting. On our construal, she knew full well what she was doing when she chose candidates with white-sounding names. She may have been helped along by her implicit biases, but she was not coerced in any way. Upon reflection, she would endorse her contribution to the committee, its procedures, and the hire in which they culminated. Her actions, assessments, and decisions, and the outcome they helped bring about are, in fact, an expression of her considered intentions. It is not clear that she would want recourse to any exculpating condition, and in any event does not satisfy either one. She is straightforwardly responsible and deserves considerable blame.

A much more interesting matter lies in what we think is an important difference between the old and new egalitarians. We described them as both having roughly the same psychological profile—the same avowed egalitarian ideals, but also the same implicit racial biases—and contributing equally to the same unjust outcome; so one might be tempted to say that they both deserve roughly the same amount of blame as well. We think this would be a mistake. Rather, we hold that the new-school egalitarian is considerably more responsible than the old, and that more blame should be directed at him. Although neither knew they had implicit racial biases, and neither would endorse those biases or their manifestation in this case, given the relevant differences in external contexts and wider social circumstances, especially the psychological research that accumulated between 1980 and 2014, the new egalitarian *could* have and *ought* to have known about this, and *could* have and *ought* to have taken appropriate steps to nullify or counteract their influence on the decision process. Since the same cannot be said of the old-school egalitarian, she does not deserve nearly as much blame as the new. Times change; excuses wear thin.

To begin developing this intuition and what lies behind it, consider some arguments that, if successful, would absolve *both* egalitarians. One way to try to do this would be by appeal to a control-based exculpating condition, expressing a form of the Bypassing Worry we mentioned previously. This does not strike us as a promising way to go. First, we reject the claim that neither egalitarian had *any* kind of control over their decisions, or the resulting hire.[12] Indeed, for any of the

[12] The question of what kind of control is required for moral responsibility is a controversial one, to say the least. For some discussion of the different types of control that philosophers have thought to be relevant, see Holroyd and Kelly (forthcoming), who also defend a two-step argument that 1) implicit biases are in fact subject to what Andy Clark (2007) calls "ecological control," and that 2) ecological control is sufficient for moral responsibility.

three individuals on the committee, it seems utterly implausible to us that any person's explicit deliberative capacities were *completely* bypassed, or that what each individual experienced as conscious reflection over the course of the hiring process was purely epiphenomenal. Unlike the kind of snap decisions made by NBA referees, evaluating and ranking resumés is a case of a slow, temporally extended decision-making process. In such cases, conscious reflection is likely to be an important and causally efficacious element of the story, and so the resulting decision is certainly not a direct, unadulterated expression of implicit bias. Rather, the contribution of this aspect of the cognitive monster to the slow, temporally extended procedure is more like an illicit thumb on the scales of deliberation, contaminating the process. Moreover, it is a corrupting influence that, if recognized, could be neutralized in the type of case featured in our thought experiment, e.g. perhaps by removing or blinding oneself to the names from the resumés beforehand, or by using one of the other techniques that research has shown to be effective. The upshot of this is that if there is a problem for holding either of the egalitarians responsible, it is not that their conscious, deliberative, or other agential capacities were completely bypassed by the psychological processes resulting in the hiring decision. Therefore, if either can be found free of fault, the most plausible way to make the case is in terms of the Epistemic Worry rather than the Bypassing Worry.

Imagine an analogous (and perhaps dispiritingly familiar) case: you have a clueless student in one of your introductory classes who fails to show up to the midterm, but afterwards pleads his case to you. He claims that it was not his fault that he was absent because he was unaware that the examination was on that day. He begs for mercy, and asks to be allowed to retake the examination; he did not mean to blow it off, he just forgot to check the syllabus, and so did not know. Cast in our terms, the student is clearly making appeal to the knowledge-based exculpating condition. Moreover, we are willing to concede that he in fact satisfies the condition. But of course, all things considered, he is responsible for missing the examination (and you would be well within your rights to deny his plea to retake it). Even though the student genuinely did not know when the examination was taking place, his mere ignorance does not excuse his absence. He should have known when the midterm was; it was his responsibility to know. Unfortunately for him, he satisfies the exception clause, too.

Importantly, it is not *everyone's* responsibility to know this fact about the midterm. Rather, the student inherits this and other responsibilities in virtue of occupying a certain social role—in this case, being a member of the introductory class. In signing up for the class, it became the student's responsibility to know certain things, such as the date of the examination and the rest of the contents of the syllabus. Particular responsibilities attach to other social roles as well. Consider a negligent doctor who fails to keep up with current medical findings and techniques, and so loses a patient who could have been saved easily by a relatively new, life-saving procedure of which she was unaware. Once again, the doctor genuinely does not know about the new life-saving procedure, but this ignorance alone does not excuse her. As a medical doctor it is her responsibility *to know certain things*, to keep abreast of the current state of medical knowledge. Not everyone is required or expected to have this knowledge, but certain people most definitely are, and if they do not, they are blamed for bad things that happen as a result of their ignorance. As in the case of the clueless student, we grant that in the scenario described, our negligent doctor satisfies the knowledge-based exculpating condition, but we maintain that she *also* satisfies the exception clause. She is not excused, and so is straightforwardly responsible for the situation, and deserves blame for her patient's death.[13]

With these examples in hand, we can say more precisely how knowledge of implicit bias affects the respective responsibility of our hiring committee egalitarians. But first note another commonality between the two: not only do they both share roughly the same psychological profile, but in being on the same hiring committee they also both occupy the same social role. Despite this, our assessments of the two diverge. While the old-school egalitarian did contribute to a morally problematic outcome, she herself is absolved of responsibility for that outcome, and bears little blame. She did not know that she harbored implicit racial biases, nor was she aware of implicit biases in general or how they were affecting her deliberations and judgments in assessing the resumés. In 1980, *no one* knew the unsettling psychological facts about implicit biases; the psychological research had not yet been done, and so today's wealth of empirical evidence simply did not exist. The old-school egalitarian is absolved in virtue of meeting the knowledge condition. She did not know about implicit biases, and could not have been expected to know about them.

The new egalitarian does not get off so easily. He has much in common with the old: he occupies the same social role as a member of the hiring committee,

[13] The negligent doctor case is similar to a case from Holly Smith's 1983 paper "Culpable ignorance." For Smith, an individual is culpably ignorant if they behave in a way that demonstrates an irresponsible willingness to take risks. We agree with Smith that there is a meaningful distinction between an individual who does not know any better and an individual who ought to have known better. In other words, for an individual to be culpable there must be a "benighting" act, in which irresponsible risk-taking occurs. But this does not mean that the benighting act is *all* that the individual is responsible for. We contend that individuals are responsible for their risk-taking *and* their later "unwitting" acts—the negligent doctor is responsible for not having kept up with her craft *and* for each subsequent patient she injures.

and contributes to the same morally problematic outcome. His deliberative process was not bypassed, nor was he coerced in any other way. He was likewise ignorant of his own implicit biases and their influence, so he satisfies the knowledge condition. But the new egalitarian also meets an exception clause, and therefore is responsible and bears more blame for the outcome than the old-school egalitarian does. This difference is a function of the external context and wider social circumstances he inhabits in virtue of being indexed to the year 2014. Today, the amount of empirical evidence collected on implicit biases is enormous, and it continues to mount. Much more is known in general, and that knowledge is much more widespread in the new egalitarian's epistemic environment than it was in the early 1980s environment of the old-school egalitarian.[14] The new egalitarian, like the old, does not know about his own implicit biases—but *unlike* the old-school egalitarian, he *should have been aware*. The differences in their wider social circumstances and informational environments are represented in the different relations each bears to the exception clause: only the new egalitarian satisfies the knowledge condition's exception clause, and thus bears considerably more blame than his old-school counterpart.

5 Taking a Step Back: Externalism and the Evolution of the Epistemic Environment

We hope readers share our initial intuition that there is something questionable about the new egalitarian when compared to his old-school counterpart, and we hope to have begun to unpack what lies behind that intuition in a way that helps clarify and strengthen the assessment it supports. In this section we will take a step back from the specifics of that case and reflect on some of the more general features of our approach, and point to some questions that it raises. While recent discoveries about a set of counterintuitive and unsettling facts of human psychology are at the forefront of our discussion, there is also something traditional about the approach we have taken to them. We understand ourselves to primarily be doing moral philosophy. We have not added to the evidence about implicit

[14] As we discuss in Section 6, facts about implicit biases are not yet a matter of common knowledge, but information about them continues to be disseminated into the wider culture. Since 1998, more than 16 million people have taken on online IAT (Brian Nosek, personal communication). Moreover, in the US, high-profile cases, such as those involving Trayvon Martin and Michael Brown, continue to bring media attention to racial bias, and implicit biases have been increasingly discussed in the popular press commentary on those cases (see e.g. <http://www.huffingtonpost.com/tag/implicit-bias> <http://www.motherjones.com/politics/2014/11/science-of-racism-prejudice> <http://www.nytimes.com/2015/01/04/upshot/the-measuring-sticks-of-racial-bias-.html?_r=0>).

bias, nor argued for a new interpretation of the extant data. But neither have we urged that taking account of those unsettling facts will require extensive revision of our moral concepts or radical overhaul of the currently entrenched practices surrounding ascription of responsibility and blame.[15] Rather, we started with typical, recognizable patterns of moral reasoning, as represented in our exculpating conditions and exception clauses, and tried to show how they can be extended to deal sensibly with a class of cases that involve implicit bias-influenced behavior. Our discussions of the mundane examples of Cate the cookie thief, the clueless student, and the negligent doctor were designed to illustrate this point.

However, the putatively uncontroversial general premises that we take as our starting point—knowledge is relevant to moral responsibility, differences in knowledge can be reflected in differences in responsibility and blameworthiness, changes in knowledge can generate changes in responsibility and blameworthiness—can lead to less banal conclusions when combined with premises inspired by the thriving research on implicit biases. What is novel about cases involving implicit bias driven behavior is the psychology, and perhaps the epistemology of that psychology. As with the clueless student and negligent doctor, we are finding fault with our new egalitarian for failing to know something that he should have known. However, unlike the first two cases, some of the facts about which the new egalitarian is ignorant are facts about himself, his own mental processes and tendencies. Moreover, the new egalitarian cannot gain the relevant knowledge about these mental states—the character of implicit biases, and that such mental states are present and operative in his own psychological apparatus—simply by introspecting. But while the mind is not here transparent to itself, the new egalitarian can become aware of them. Knowledge of them just has to come via less direct, often external pathways (perhaps by taking an IAT online); in this sense, the epistemology of one's own implicit biases is non-Cartesian. Moreover, empirical research suggests that implicit biases cannot be well controlled via direct or immediate exercise of willpower. Rather, successfully curbing the influence of one's implicit biases will first require the acquisition of more and different knowledge from without. For not only does an individual need to know that she has implicit biases before she can even try to exert control over them, but doing so consistently

[15] Compare with Doris (2015), who considers a much larger array of counterintuitive and creepy facts revealed by recent psychological research, and argues that these present us with a dilemma. We must either radically revamp the conception of agency found in much of the philosophical literature, which gives pride of place to reflection and conscious deliberation, or, if we hold onto the reflectivist conception, we will be driven towards skepticism about persons, and the conclusion that genuine episodes of agency actually occur *much* less frequently than previously thought.

and effectively will also require a special kind of knowledge—specifically, knowledge of and facility with the kind of techniques and methods that are being shown to be effective by the empirical research on the malleability of implicit bias (see fn. 10).

Our line of reasoning dovetails with other anti-Cartesian trends in the philosophy of mind and cognitive science that often fall under the banner of externalism. There are many forms of externalism, but the common thread is an insistence that the boundaries of an individual's skin and skull are relatively unimportant when it comes to the nature and content of her mind, and the bases of her judgments and behavior. Some externalists have famously claimed that the content of mental states is in part determined by factors outside of the head, while others have gone even farther, arguing that mental states and cognitive processes themselves can extend beyond the borders of a person's physical, organic body.[16] Similarly, in our thought experiment, the most important difference between our old and new egalitarians is not something within the boundaries of their skin, but is rather in the wider social and cultural circumstances in which each is situated, as captured by the years to which each is respectively indexed. Indeed, one implication of the intuition we are pumping is that not all of the knowledge relevant to moral responsibility and exculpation need be "in the head" of the agent whose actions are being evaluated.

Even this externalist aspect of our view is somewhat traditional—it appears to be true of the kind of reasoning that applies to cases like the clueless student and the negligent doctor. In the second case, for instance, there was information about a new, life-saving procedure in her cultural environment, but she was just not aware of it; it was in the journals, clinics, and other medical practitioners' heads,

[16] See Putnam (1975), Burge (1979), and Fodor (1987, 1994) for defenses of what has become known as passive or semantic externalism, Clark and Chalmers (1998) for the initial statement of the extended mind thesis and what has become known as active or vehicle externalism, and Dennett (2003), Clark (2007), Shapiro (2007), and Ismael (2007) for the development of similar ideas. Other approaches that have an externalist flavor emphasize different aspects of the extrabodily environment and the different roles they can play in human psychology, often creating new terminology to talk about them. For instance, see Doris (1998, 2002) for discussion of the underappreciated role of external situational factors, both physical and social, in driving behavior, and Merritt (2000) for a development of the core idea that emphasizes how properly structured environments can make a sustaining social contribution to ethical behavior. Sterelny (2003, 2012) extends the conceptual resources of niche construction theory to show how humans actively engineer the informational niches in which they live, learn, and raise children, and argues that this deliberate organizing of their own epistemic environment is a key factor in explaining human behavior and evolution. Defenders of gene culture coevolutionary theory stress the importance of social learning and the accumulation of cultural information, which is often contained in brains, but can also be manifest in behavior, realized in artifacts, written in books, and so on. The name of the theory indicates that it construes the repository of cultural information as an inheritance system that operates in tandem and interacts with the genetic inheritance system, and whose contents are subject to analogous kinds of selective pressures (see e.g. Richerson and Boyd (2005); Boyd and Richerson (2005); Henrich (2011)).

but it was not in *her* head. Had that information been absent not just from her head, but from the doctor's epistemic environment in general, however—if the new life-saving procedure had not yet been developed—then she would not have been to blame her for patient's death. Again, we maintain that similar reasoning applies to the two egalitarians in our thought experiment.[17] Differences in knowledge can produce differences in responsibility and blameworthiness, and changes in knowledge can generate changes in responsibility and blameworthiness—even when those changes are in the informational content of the cultural and epistemic environment, rather than in the head of the agent being evaluated.

Perhaps somewhat oddly, the case involving implicit bias leads us to the position that not all of the knowledge relevant to moral responsibility and exculpation need be in the head of the agent whose actions are being evaluated, even when the *subject matter* of that knowledge is, in fact, in her very own head! One can take the slight whiff of paradox out of this if one considers a parallel case of knowing one's own cholesterol level or blood pressure. You might think that today, as a well-informed adult, part of taking responsibility for your own health and well-being is keeping track of these physiological features of your own body, and taking steps to keep them at acceptable levels.[18] But of course, this could not have been the case for, say, someone living in Shakespeare's day. It is dependent on advances in medical knowledge—a whole slew of discoveries made only in the last few centuries. No one innately grasps truths about cholesterol or blood pressure, nor knows intuitively that these are important indicators of health and should be monitored. Nor can anyone introspectively discern her own cholesterol level; you have to look without to gain that knowledge about yourself. For most of us this involves going to a doctor, who will use technologically sophisticated instruments to take measurements whose meaning she will report back to you. Moreover, no one can directly control her own blood pressure, or

[17] While our approach focuses on individual responsibility, we are also alert to the fact that the externalist orientation can inspire similar arguments about the responsibility that institutions have to take steps against bias and prejudice, and to structure the institutional environment of those individuals who operate within them. We briefly discuss this idea in Section 6, and also think that a more sustained look at the interaction of collective and individual responsibility is a worthwhile undertaking. Construing the project (of attempting to raise awareness, alter social norms, and reform institutions in the relevant ways) as a specific attempt at guided cultural evolution could provide useful insight into how best approach the task.

[18] Even today, this is only the case in some cultures, those with the readily available technology and properly disseminated medical information. The variability in cultures, and the resultant differences in what "taking responsibility for your own health and well-being" amounts to, is compatible with the externalist orientation we favor, and the idea that differences in the informational environment and moral ecology can yield different kinds of responsibility on the individuals who inhabit them.

bring about an immediate and sustained change in it by direct act of will. To effectively control our blood pressure, most of us need to learn about and use the more roundabout, external methods that have been empirically verified. You will have to take slower, less direct steps to bring about the sought after change, even though the change is internal, and in your own physiological makeup and functioning. In these respects, our view is: as with cholesterol and blood pressure, so with implicit bias.[19]

6 A Glimpse Ahead: Objections and Open Questions

A series of further questions can be asked about comparative levels of responsibility. There are important questions not just about who is responsible, but also about how much responsibility we are ascribing to the different members of the hiring committee, and, for the two we did deem responsible, how blameworthy each is, respectively, and what form of punishment would be most appropriate (not to mention effective).[20] For the purposes of this paper, we have largely left our discussion at an intuitive level, and hope that, as such, it can be plugged into different, more sophisticated ways of answering such questions, and making the attendant notions more precise. For now, we will comment on a point that has been made to us a number of times: namely, that our approach seems to have assumed a qualitative notion of responsibility, but one that is comparative and graded as well; i.e. we say that the new egalitarian is less responsible than the explicit racist. Whether or not sense can be made of this way of construing responsibility ascription is an interesting question, but even if it turns out to be unworkable, we do not think it would threaten our main point. As far as we can see, everything we have said is also compatible with a view according to which responsibility is not graded, but rather a more binary, all-or-nothing notion—the

[19] Using Clark's (2007) useful terminology, we could make this point by saying that one needs to use *ecological control* to effectively influence one's own blood pressure and cholesterol levels; also see Holroyd and Kelly (forthcoming) for discussion of ecological control and implicit bias.

[20] While some alarmists worry that holding each other responsible for behaviors influenced by implicit bias is likely to provoke strongly negative, counterproductive reactions, and perhaps even an increase in biased tendencies, this is not always what happens. It turns out that some forms of finding and addressing fault can ultimately help to bring about more positive results. For example, the effects of interpersonal confrontation are much less straightforward than might be expected. In a series of papers, Alexander Czopp and colleagues have explored different aspects of the phenomenon, showing that confronting a person who expresses bias reduces that person's prejudicial behavior (Czopp et al., 2006). Even more intriguing, they also found that failing to confront prejudicial behavior that one witnesses can lead to an increase in one's own bias (Rasinski et al., 2013). Moreover, there could also be important individual differences here, with different techniques more likely to work on different individuals; for instance, in their reaction to and susceptibility to guilt.

assessment of an action in light of the relevant exculpating conditions will find the agent to be either responsible or not. The need for a more fine-grained spectrum of distinctions can still be met by graded notions of blame and blameworthiness, which are brought to bear only on actions to which the binary notion of responsibility applies. In these terms, the explicit racist and the new egalitarian are both equally responsible, but the former bears more blame than the latter, which might be reflected in a more severe form of punishment. Whether and how any of these qualitative, comparative notions can be made more precise by being interpreted in a quantitative framework are intriguing questions, but ones that we cannot fully address here (though see Kagan, 2012, for such a framework).

Another objection we have heard in presenting this material accepts our argument that the new egalitarian is responsible and blameworthy, but holds that we are being too easy on the old-school egalitarian. She is also responsible for the unjust job search because she too should have known better. Bias, prejudice, and discrimination have been with us for a long time, goes the objection, and people have known about it for just as long. An observer of the human scene astute enough to appreciate prejudicial behavior for what it is would also notice that such behavior can be more or less overt. If the old-school egalitarian genuinely holds the values she professes to hold, she should have been alert to the possibility of prejudice covertly influencing her participation in the hiring decision, and taken steps to prevent it. In effect, this objection accuses us of overplaying the relevant differences between 1980 and 2014.[21]

We have no doubt that bias and discrimination have been a part of the human condition as long as there has been one, and that enlightened individuals have been able to recognize it and many of its forms as such. However, we also think that the advances in empirical psychology make a collective difference, and the sheer amount of evidence on implicit biases constitutes a crucial one. As captured by Bargh's invocation of the cognitive monster, there is a natural suspicion that these advances and evidence make a *moral* difference, that they have moral significance; indeed, we take ourselves to be showing one way to flesh out that suspicion. Someone living prior to 1980 could be sufficiently observant to discern that something funny and potentially biased was going on in the kind of hiring process we imagined. But she certainly could not have known, let alone been responsible for having known, anything terribly specific about the kinds of

[21] We are thankful to Alex Madva for pushing us to think about this; see Madva (this volume) and Brownstein and Madva (2012).

psychological processes that were driving it.[22] That implicit biases are not easily visible to folk psychology and the folk morality that depends on it, and that the details about them are so counterintuitive—they are so widespread, and can automatically influence a variety of behaviors and judgments, can co-exist with considered views to the contrary, are opaque to introspection and resistant to common forms of control—only strengthens our position on this. But now we do know, and that matters. There have been specific changes in the collective knowledge about *these* specific psychological processes, and changes in the collective knowledge of the possibility and likelihood that *these very psychological processes* will influence certain outcomes in *these very ways*, and that their influence can be mitigated with *these* kinds of methods but not *these*. More and more specific things are known about biases and the sources of prejudiced behavior now, and that, we maintain, should be reflected in shifts in how people should be held responsible for them. As our collective informational environment evolves in this way, our moral ecology should evolve with it.[23]

Of course, as impressive as the accumulated knowledge about implicit biases is, it has still not risen to the level of *common knowledge*, and it probably will not at any time in the immediate future. Different members of the population certainly have different levels of familiarity with the research on these unsettling facts. Judging from our own experience, people with the psychological profile of our new egalitarian (explicitly egalitarian, vaguely aware of the existence of implicit bias in general, but unaware of their own) remain common, and the percentage of people who have not heard of implicit bias at all is probably still quite high. Considering this observation in light of our evaluation of the members of the

[22] Another set of unsettling facts highlighted by recent empirical psychological work centers on the normal human tendency to confabulate, and a susceptibility for believing self-serving rationalizations about what motivates our behaviors and judgments. Given this, we also think that asking someone like our old-school egalitarian to have enough clarity and self-awareness to see through her own rationalizations and glean what was actually happening in 1980, without benefit the empirical record, or even the clear idea of a mental state with a profile of characteristics like that of implicit biases, would have been asking unreasonably much of her, to put it lightly.

[23] We also see a fellow traveler in Miranda Fricker, and are sympathetic to much of her discussion about the relativism of blame (2010). She notes that "sometimes an agent may be living at a time of transition," and that when we in the present are making retrospective judgments about agents who lived during such transitional periods in the past, there is a "form of critical moral judgment" we can use on those who are "slow to pick up on" the period's "dawning moral–epistemic innovation" that she calls *moral–epistemic disappointment*. Our old-school egalitarian seems to fit this bill quite well; he is on the cusp of an important transition in our understanding of the psychology of prejudice. As Madva emphasizes, suspicion that there is more to prejudice and bias than the full-blown explicit kind has been around for a long time, but as we emphasize, the accumulation of experimental evidence and detailed understanding of the specific character and operation of implicit biases has been a recent development. As such, we can say that one appropriate attitude we might take to our old-school egalitarian is that of moral–epistemic disappointment.

hiring committee can prompt the worry that our approach does something like penalize the wrong people: those admirably curious people who stay well informed enough to become alert to the existence of implicit biases become responsible for their own, and thereby open themselves up to blame if they fail to take appropriate measures to deal with them.

We have three things to say in response. First, our position is that now that knowledge about implicit bias is part of the informational environment, certain people can inherit the *responsibility to know* about it, and to know about and deal with their own. However, we are not suggesting that it is equally everyone's responsibility—not even everyone in 2014—to know about and take precautions again implicit biases. Responsibility to know about implicit bias, like responsibility to know about the contents of a syllabus or about advances in medical knowledge, is not (yet) distributed equally across all people in a population, nor across all agents in virtue of being agents. Rather, we think that as research about these unsettling facts and their unsavory influence mounts, the responsibility to take preventive measures against them accrues first, or at least more quickly and disproportionately, to occupiers of specific social roles. These include those involved in hiring decisions, obviously, but also teachers, social workers, and those in other "gatekeeper" positions whose activities can have the most amplified effects on various institutions and population level outcomes. Exactly which social roles this responsibility falls upon, and at what point during the accumulation of research on implicit bias and the evolution of the wider informational environment, is an important and interesting question. We take ourselves to have made progress in raising the question in this form, but do not yet have an answer to it.

Second, bracketing the idea of a role specific responsibility to know, we can better address the worry about "penalizing" the wrong people in virtue of what they actually do or do not know about implicit biases (rather than what they should know). In short, we are willing to bite the bullet on this point. Those who harbor implicit biases but have no knowledge of them (and do not occupy one of relevant social roles) can satisfy the knowledge-based exculpating condition without also meeting the exception clause. In such cases, they should not be held responsible for their actions, and are not blameworthy. In this sense, those who *are* in the know are indeed the first to inherit the "burden" of responsibility for their own implicit biases. This does not strike us as particularly surprising, or wrong. Someone has to lead the way, and the kinds of culturally sophisticated, ethically motivated, self-aware people who are likely to hear about implicit biases and take the time to find out if they harbor any themselves are exactly the type of people we would expect to be willing to act as examples for the rest.

This, however, brings us to our final point, which is just to emphasize that the situation continues to evolve. Indeed, those interested in progress need not just passively watch it happen, but can channel their efforts in informed ways to help guide the process. Of course, they can act as role models and vocal advocates in traditional ways. But the perspective we favor also shows the value of disseminating the psychological research, publicizing it, and making it part of the common knowledge of the population: raising awareness can raise the standards of responsibility as well. Activists can encourage people to take implicit association tests, and press institutions to require certain members to do so. As a greater segment of the population comes to know about their implicit biases, more and more people can be held responsible for dealing with them. As psychologists learn more, those interested in change can keep packaging the relevant parts of that research in ways that make it easiest for the uninitiated to understand, and continue pumping it out into the wider social environment. In this way we can continue in actively constructing the larger informational environment. In shaping our epistemic niche, we are also guiding evolution of the moral ecology that we all inhabit. One ideal to aim at is in engineering an environment in which everyone is expected to know about and deal with their own implicit biases—not because they occupy any specific social role, but because it is a responsibility one has simply in virtue of being a person.

7 In Place of a Conclusion

We certainly realize that there remains work to be done, but maintain that the ongoing project will benefit from being informed by the broadly externalist orientation we have been urging. Our main specific aim in this paper is to show that there are clear cases in which a person should be held responsible for behaviors influenced by implicit bias, even if she does not know she has the implicit bias. We have defended this claim by making explicit and illustrating patterns of moral reasoning that typically accompany currently existing norms and practices, and showing how they can be extended to deal with certain cases of behavior driven by these kinds of counterintuitive and unsettling psychological processes. The approach we have taken draws inspiration from the growing family of views emerging in philosophy of mind and the cognitive sciences united by their insistence that what is outside the head can be just as important as what is inside it. We have used all of these resources to map out one way that moral significance can be given to the accumulating empirical evidence about implicit biases.

We want to bring the discussion to a close by putting a positive twist on what could otherwise seem like a depressing or vindictive endeavor, as if our motivation was primarily to find fault, identify targets for blame, and perhaps implying that they deserve punishment—that we are mainly out to get someone. To be sure, fully appreciating the potential power of implicit biases and the kinds of outcomes they help produce and sustain can be disheartening. But another, more uplifting way to look at the significance of the mounting empirical evidence and the consequent changes in the moral ecology that it can spark is in terms of the increases of freedom they might purchase. From this point of view, the trajectory of the changes in the larger epistemic environment on which we have been focusing (e.g. advances in empirical psychology from 1980 to 2014, and hopefully on into the future) have been largely in the direction of improvement, and provide reason for optimism. Not only do those changes constitute genuine *empirical progress*, but we hope to have begun to sketch out how they can be translated into *moral progress* as well. In better understanding, publicizing, and ultimately using what is being discovered about implicit biases and how to most effectively mitigate their influence, we can take more responsibility for ourselves, buying both freedom *from* the unwanted influence of these unsettling mental processes, and the freedom *to* be able to act in ways that we wish—ways that more closely fit with our considered ideals. Understanding our own minds can help us to take responsibility for ourselves more fully, and more often. Understood properly, empirical discoveries about implicit biases and other unsettling psychological facts can eventually, and perhaps unexpectedly, show us how to increase our agency, do better things, and be better people.

Acknowledgments

We would like to thank the following people for useful questions and feedback on earlier presentations and drafts of this material: Michael Brownstein, Luc Faucher, Jules Holroyd, Alex Madva, Chandra Sripada, the members of the Purdue Psychology Department Brownbag talk series, Jennifer Saul and the attendees of the Implicit Bias and Philosophy workshops held at the University of Sheffield, Andreas de Block and the members of the Faculty of Business and Economics at KU Leuven, members of the Moral Psychology Research Group, Cathrine Felix and the members of the Filosofiska Föreningen at Lund University, and the audience members at the 2013 SSPP symposium on Implicit Bias and Moral Responsibility.

References

Banaji, M. and Greenwald, A. G. (2013). *Blindspot: Hidden Biases of Good People*. New York, NY: Delacorte Press.

Bargh, J. A. (1999). "The cognitive monster: The case against controllability of automatic stereotype effects." In Chaiken, S. and Trope, Y. (eds.), *Dual Process Theories in Social Psychology*. New York: Guilford Press: 361–82.

Bertrand, M. and Mullainathan, S. (2004). “Are Emily and Greg more employable than Lakisha and Jamal?: A field experiment on labor market and discrimination.” *American Economic Review* 94(4): 991–1013.

Blair, I., Steiner, J., and Havranek, E. (2011). “Unconscious (implicit) bias and health disparities.” *The Permanente Journal* 15(2): 71–8.

Boyd, R. and Richerson, P. (2005). *The Origin and Evolution of Cultures*. New York: Oxford University Press.

Brownstein, M. and Madva, A. (2012). “The normativity of automaticity.” *Mind and Language* 27(4): 410–34.

Brownstein, M. and Saul, J. (eds.) (2015). *Implicit Bias and Philosophy. Volume I: Metaphysics and Epistemology*. Oxford: Oxford University Press.

Burge, T. (1979). “Individualism and the mental.” In French, P. A., Uehling, T. E., and Wettstein, H. K. (eds.), *Midwest Studies in Philosophy*, vol. 4: Minneapolis: MN: University of Minnesota Press: 73–121.

Carr, P. B. and Steele, C. M. (2010). “Stereotype threat affects financial decision making.” *Psychological Science* 21(10): 1411–16.

Clark, A. (2007). “Soft selves and ecological control.” In Spurrett, D., Ross, D., Kincaid, H., and Stephens, L. (eds.), *Distributed Cognition and the Will*. Cambridge, MA: MIT Press.

Clark, A. and Chalmers, D. J. (1998). “The extended mind.” *Analysis* 58: 7–19.

Czopp, A., Monteith, M., and Mark, A. (2006). “Standing up for change: Reducing bias through interpersonal confrontation.” *Journal of Personality and Social Psychology* 90(5): 784–803.

Dennett, D. (2003). *Freedom Evolves*. Penguin Books.

Dennett, D. (2013). *Intuition Pumps and Other Tools for Thinking*. New York, NY: W. W. Norton.

Doris, J. (1998). “Persons, situations, and virtue ethics.” *Noûs* 32: 504–30.

Doris, J. (2002). *Lack of Character: Personality and Moral Behavior*. New York, NY: Cambridge University Press.

Doris, J. (2015). *Talking to Ourselves: Reflection, Ignorance, and Agency*. Oxford University Press.

Fischer, J. and Tognazzini, N. (2009). “The truth about tracing.” *Noûs* 43: 531–56.

Fodor, J. (1987). *Psychosemantics*. Cambridge, MA: MIT Press.

Fodor, J. (1994). *The Elm and the Expert*. Cambridge, MA: MIT Press.

Fricker, M. (2010). “The relativism of blame and William’s relativism of distance.” *Aristotelian Society Supplementary Volume* 84(1): 151–77.

Haidt, J. (2006). *The Happiness Hypothesis: Finding Modern Truth in Ancient Wisdom*. New York, NY: Basic Books.

Haney-López, I. (2000). “Institutional racism: Judicial conduct and a new theory of racial discrimination.” *The Yale Law Journal* 109(8): 1717–885.

Henrich, J. (2011). “A cultural species: How culture drove human evolution.” *Psychological Science Agenda*. Science Brief: <http://www.apa.org/science/about/psa/2011/11/human-evolution.aspx>.

Henrich, J., Heine, S., and Norenzayan, A. (2010). “The weirdest people in the world.” *Behavioral and Brain Sciences* 33: 61–135.

Hirstein, W. (2005). *Brain Fiction: Self-Deception and the Riddle of Confabulation.* Cambridge, MA, and London: MIT Press.

Holroyd, J. (2012). "Responsibility for implicit bias." *Journal of Social Philosophy* 43(3): 274–306.

Holroyd, J. and Kelly, D. (forthcoming). "Implicit responsibility character and control." In Webber, J. and Masala, A. (eds.), *From Personality to Virtue.* Oxford: Oxford University Press.

Ismael, J. (2007). *The Situated Self.* Oxford: Oxford University Press.

Kagan, S. (2012). *The Geometry of Desert.* New York, NY: Oxford University Press.

Kawakami, K., Dovidio, J. F., and van Kamp, S. (2007). "The impact of näıve theories related to strategies to reduce biases and correction processes on the application of stereotypes." *Group Processes and Intergroup Relation* 10: 139–56.

Kelly, D. and Roedder, E. (2008). "'Racial cognition and the ethics of implicit bias." *Philosophy Compass* 3(3): 522–40. doi:10.1111/j.1747-9991.2008.00138.x

Machery, E., Faucher, L., and Kelly, D. (2010). "On the alleged inadequacies of psychological explanations of racism." *The Monist* 93(2): 228–55.

Madva, A. (ms.). "Biased against de-biasing: On the role of (institutionally sponsored) self-transformation in the struggle against prejudice."

Merritt, M. (2000). "Virtue ethics and situationist personality psychology." *Ethical Theory and Moral Practice* 3: 365–83.

Nahmias, E. (2010). "Scientific challenges to free will." In O'Connor, T. and Sandis, C. (eds.), *A Companion to the Philosophy of Action.* New York, NY: Wiley–Blackwell.

Nisbett, R. E. and Wilson, T. D. (1977). "Telling more than we can know: verbal reports on mental processes." *Psychological Review* 84(3): 231–59.

Payne, B. K. (2005). "Conceptualizing control in social cognition: The role of automatic and controlled processes in misperceiving a weapon." *Journal of Personality Social Psychology* 81: 181–92.

Payne, B. K. (2006). "Weapon bias: Split-second decisions and unintended stereotyping." *Current Directions in Psychological Science* 15: 287–91.

Price, J. and Wolfers, J. (2010). "Racial discrimination among NBA referees." *Quarterly Journal of Economics* 125(4): 1859–87.

Putnam, Hilary (1975). "The meaning of meaning." In *Philosophical Papers, Vol. II: Mind, Language, and Reality.* Cambridge: Cambridge University Press: 215–71.

Rasinski, H., Geers, A., and Czopp, A. (2013). '"I guess what he said wasn't that bad': Dissonance in nonconfronting targets of prejudice." *Social Psychology Bulletin* 39(7): 856–69.

Richerson, P. and Boyd, R. (2005). *Not by Genes Alone.* Chicago, IL: University of Chicago Press.

Roskies, A. (2006). "Neuroscientific challenges to free will and responsibility." *Trends in Cognitive Science* 10(9): 419–23.

Saul, J. (2013.) "Implicit bias, stereotype threat and women in philosophy." In Jenkins, F. and Hutchison, K. (eds.), *Women in Philosophy: What Needs to Change*? New York, NY: Oxford University Press: 39–60.

Shapiro, L. (2007). "The embodied cognition research programme." *Philosophy Compass* 2(2): 338–46.

Shoemaker, D. (2011). "Attributability, answerability, and accountability: Toward a wider theory of moral responsibility.' *Ethics* 121(3): 602–32.

Smith, A. (2012). "Attributability, answerability, and accountability: In defense of a unified account." *Ethics* 122(3): 575–89.

Smith, H. (1983). "Culpable ignorance." *Philosophical Review* 92(4): 543–71.

Smith, H. (2011). "Non-tracing cases of culpable ignorance." *Criminal Law and Philosophy* 5(2): 115–46.

Sommers, T. (2011). *Relative Justice: Cultural Diversity, Free Will, and Moral Responsibility*. Princeton, NJ: Princeton University Press.

Sripada, C. (forthcoming). "Self-expression: A deep self theory of moral responsibility." *Philosophical Studies*.

Staats, C. and Patton, C. (2013). *State of the Science: Implicit Bias Review 2013*. Columbus, OH: The Kirwan Institute.

Sterelny, K. (2003). *Thought in a Hostile World*. New York, NY: Blackwell.

Sterelny, K. (2012). *The Evolved Apprentice: How Evolution Made Humans Unique*. Cambridge, MA: MIT Press.

Tavris, C. and Aronson, E. (2007). *Mistakes Were Made (But Not By Me)*. New York, NY: Harcourt.

Vargas, M. (2005). "The trouble with tracing." *Midwest Studies in Philosophy* 29(1): 269–91.

Vargas, M. (2013). *Building Better Beings: A Theory of Moral Responsibility*. Oxford: Oxford University Press.

Watson, G. (1996). "Two faces of responsibility." *Philosophical Topics* 24(2): 227–48.

Wilson, T. D. (2002). *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Cambridge, MA: Harvard University Press.

1.2

Alienation and Responsibility

Joshua Glasgow

We have troubling implicit biases. Without being aware of it, we more readily associate white people with good qualities and black people with bad qualities, we are more likely to mistake a harmless object for a gun if held by a black person as opposed to a white person, we are less likely to associate a person of Asian descent than one of European descent with being an American, and we are less likely to select for an interview an applicant with a woman's name than one with a man's name.[1] When testing reveals these implicit biases, many people, of course, wholeheartedly and unequivocally disavow them. They do not *deny* that they have the biases (though some do that, too); rather, they sincerely and truly claim that their biases do not represent who they *really* are. In part because of the discordance between their biases and their self-image and values, many who have implicit biases consider themselves *guilty*, not just embarrassed but *worthy* of embarrassment. They consider themselves *responsible* for these disavowed attitudes.

For one representative example of this common response, consider the following statement from Mathieu Marion, taking responsibility for the journal he co-edits failing to include any women authors in its fiftieth anniversary issue, which republished twelve articles from its archive:

> [T]he whole thing is a terrible oversight on my part, and I have no excuse whatsoever for this to have happened, simply because there is *no* excuse. I should add that you will find my signature on the petition for the Gendered Conference Campaign and that makes my shame and embarrassment all the more vivid. Actually, what I just described as an oversight may very well be understood as my having not entirely shaken the sort of implicit bias that is prevalent in philosophy, there is no other explanation, as there is no evading the responsibility. Therefore, I can only apologize . . . in the most sincerely felt way

[1] Respectively: Dasgupta (2004); Payne (2006); Devos and Banaji (2005); Steinpreis et al. (1999).

> and beg for forgiveness for having thus harmfully misrepresented not only the true state of the discipline, but also the fact that many women have published first-rate papers in the pages of *Dialogue* through the years that could have been included in this special issue.[2]

This comment combines attribution of responsibility for implicitly biased action with (implied) disavowal of both the bias and the action. But though this is a common (which is not to say universal) reaction to implicit bias, it gives rise to a puzzle. When we truly and wholeheartedly disavow such attitudes, why should we feel guilty when we find ourselves with them, any more than we should feel guilty about our explicitly sexist uncle, whose sexism we also wholeheartedly disavow? He represents our true attitudes no less than our own unconscious biases do—by hypothesis, each is sincerely, truly, and wholeheartedly disavowed. So why do we feel guilty about the one and not the other? Why do we consider ourselves responsible for such attitudes if they are wholly alienated from our selves?

In what follows, I want to do two things. First, I try to solve this puzzle by articulating one plausible way in which we can be responsible for biases from which we are alienated (a concept about which I will say more shortly). But, second, there are more general issues in play here about responsibility under conditions of alienation. These issues, I will argue, suggest that a broader cluster of common intuitions presents a problem for Real Self Theories of responsibility—theories in which alienation figures so prominently. This problem is best fixed, I argue, by the same theory of responsibility that best fits our judgments about alienated implicit attitudes. The bigger project, then, is to make sense of certain judgments not only of responsibility for alienated bias, but of responsibility for alienated actions and attitudes in general.

I pursue these questions in a particular way. First, I want to take seriously the judgment that the disavowed attitudes and consequent behaviors in question can be morally problematic.[3] (Although implicit attitudes and consequent behaviors seem to differ from each other in terms of both susceptibility to control and self-awareness (Hall and Payne 2010), these dimensions of responsibility are orthogonal to the dimensions probed here—particularly alienation—and I will cover in one account both the attitudes and any behaviors stemming from them.) Some will think that if the attitude is disavowed, then especially if it is also mostly out of

[2] <http://feministphilosophers.wordpress.com/2012/11/04/dialogue-co-editor-apologizes/>.

[3] Whether implicit associations rise to the level of *attitudes* is an open question, the answer to which in part depends on how "attitude" functions as a term of art. I will mostly lump them under the banner of attitudes merely for notational simplicity. Those who do not like that lump can substitute "attitudes and associations" when I use "attitudes."

our control, we are not responsible for it. On this line of thought, disavowed bias is still *bad*, of course, but it is not something for which we are *morally blameworthy* (Glasgow, 2009: 79; Kelly and Roedder, 2008). But this approach fails to capture everything that many want to say about unwilling racism and other discriminatory attitudes.[4] As illustrated by Marion's comment, many think that their implicit biases are truly morally bankrupt—not just bad, but attributable to them as a moral failing. They do not just experience regret; they judge themselves responsible. I want to account for these judgments of responsibility. Accordingly, here I take the attribution of moral responsibility for disavowed biases to be a *datum* to be explained, rather than a *hypothesis* to be justified. Obviously, it may turn out that these attributions just reflect some sort of irrational guilt. Or it may turn out that they reflect guilt that one was involved in something bad without really judging that one is responsible for the bad turn of events. But the first step in sorting out whether common judgments of responsibility like the one Marion endorses are or are not well-founded, is to identify whether they are coherently explainable. We do not yet have the needed explanation. I hope to provide a plausible one in what follows.[5]

Second, since I am interested in exploring what resources are available to coherently explain the judgment of moral responsibility, it will not do to independently identify the true theory of responsibility for attitudes and then crank out an algorithmic application to a case like this. Instead, I use common judgments about disavowed bias, along with judgments about other attitudes and

[4] Cf. Adams, 1985: 21.

[5] Kelly and Roedder (2008: 532) question whether implicit bias can be disavowed to the point of being truly alienated from the agent, but they suggest that if it were, the agent might not be blameworthy for it. My aim here is to see what follows if we take seriously the intuitions that run contrary to this suggestion. Similarly, I resist Levy's (2014) argument that we can be directly responsible for explicit attitudes but not implicit attitudes. His main premise for that argument is that only explicit attitudes can generate unified agency of the right sort (and, in fact, implicit attitudes frequently undermine the kind of agency in question). This seems to me to be the basis of an argument that explicit attitudes are a precondition for responsibility. Nevertheless, once one has what it takes to be a responsible agent, other elements of one's psychology might be fair game for moral assessment; what follows takes seriously the intuition that we can have implicit bias attributed to us for responsibility. For debate about whether we should be excused from responsibility for implicit bias because, variously, it is out of our control, we are unaware of it, its presence in ourselves is not our fault, it is automatic, and it is not responsive to reasons, see Cameron et al. (2010); Hall and Payne (2010); Holroyd (2012); Kelly and Roedder (2008); Levy (2014); Machery, Faucher, and Kelly (2010); Saul (2013); and other chapters in Part 2 of this volume. For a critique of the general claim that we are not responsible for attitudes because of lack of control, unawareness, and so on, see Adams (1985). Importantly for our purposes, even those, such as Hall and Payne, who emphasize the degree to which we have awareness of, and control over, our biases also acknowledge that our biases surprise us and that we often do not take them to represent who we really are, in some sense of that phrase.

actions, to help construct a theory of responsibility that can make sense of them. If those judgments are sufficiently powerful, they will provide one motivation for accepting that newly constructed theory.

The project therefore has a distinctive character. The present agenda is to make sense of a cluster of common intuitions that are not easy to make sense of with principles commonly taken to govern responsibility. Therefore I am not very interested in defending the possible-case intuitions upon which I will draw. These intuitions are not universal—few are, of course. But they are common enough and credible enough that it is worth seeing if we can accommodate them. After all, philosophical analysis trades on such intuitions, so while they may have to be abandoned in the end, it is worth trying to find a configuration of norms that can render them coherent and intelligible. And since these intuitions seem to be very difficult to accommodate, we will have made some progress if we can find a thesis that will accommodate them all.

How can we find such accommodation? The thesis put forth here holds that, whatever else we want to say about responsibility, whether alienation from an otherwise blameworthy attitude or action exculpates depends on the particular content of that attitude or action—what the attitude or action (or omission) *is*—and that really content matters only as a proxy: what the attitude or action *is* matters ultimately because of the *harm* it causes or constitutes. This is a pretty radical view. So absent a more adequate theory, we wind up with a difficult choice: to accept an odd theory or to reject a set of judgments that are, at least to many, intuitively powerful.

1 Alienation

One venerable thread of thought, Real Self Theory, says this: if an attitude that I harbor or an action that I perform does not represent me, that is, if it does not express the part of me that is my responsible, agential self—a phenomenon for which I here use the label *alienation*—then my responsibility for it is significantly diminished. Strong versions of this view say that alienation diminishes to the fullest extent, generating complete exculpation; weaker versions say that alienation only partly exculpates. (I use "exculpation" to refer to the significant diminishment of responsibility; it can, but need not, go all the way to total cancellation of responsibility.) But either way, according to this theory, full responsibility for one's action requires that the action or the motivation for it expresses one's "real" or "deep" self, in the current vernacular.

Real Self Theorists disagree with one another as to what exactly counts as one's real self. For the early Harry Frankfurt (1971), one's real self is behind one's action when there is a certain kind of harmony among both the hierarchy of one's desires and one's action. Gary Watson's (1975) original alternative to this desiderative view suggests that responsible action is motivated by desires that comport with one's values, as opposed to mere appetites. Throughout the years, Frankfurt, Watson, and a number of others have raised problems for, and proposed other variations on, Real Self Theory. Some more recent iterations get away from the appetite/reason framework entirely. For example, Nomy Arpaly and Timothy Schroeder (1999) tether responsibility not to whether the action represents some particular part of the agent's self, like desires or values, but to whether, whatever psychological element is the source of the action, that factor is *well integrated* with the agent's personality.[6] And recently the Real Self view has even been extended beyond actions to cover attitudes, which tantalizingly could be construed broadly enough to cover implicit racial or gender associations: following T. M. Scanlon (1998: 277–94), Angela Smith (2005, 2008) holds that we are responsible for our attitudes just when they reflect our evaluative commitments in such a way that we can (in principle) be asked to justify the commitment in question.

Despite these and other important disagreements among Real Self Theorists, they all share the common stance that moral responsibility for an action or attitude is a matter of that action or attitude representing one's real self. If the act or attitude is somehow alienated from the real self—it is disassociated, externalized, disidentified with—then the agent's responsibility for the act or attitude is significantly diminished.

This view does a very good job of explaining some judgments. Frankfurt points out that it seems to be the best explanation of the common judgment that the willing addict—the one who embraces his desire to take the drug—is responsible for his addiction and its attendant behaviors, while we think that the unwilling addict—who wholeheartedly rejects his addiction but finds himself in its grip nonetheless—has his responsibility substantially diminished. Similarly, it highlights that the kleptomaniac who rejects, hates, and thoroughly disavows her desire to steal has her real self overcome when she nonetheless succumbs to the desire to steal, and for this reason she is less responsible than the willing thief.

It is important to note that nothing particularly crucial to our purposes hangs on the fact that addiction and "mania" are present in these cases. (Such

[6] For a similar view, see Smith (2011). Arpaly and Schroeder save the label "real self theory" for folks like Frankfurt in order to set it up as a foil for their "whole self theory." My including them as Real Self Theorists is merely notational: the difference between the various theories is just a disagreement over what constitutes one's responsibility-bearing, "real" self.

compulsions may be imported in the first place only because such cases vividly resonate with us as unambiguous instances of full alienation.) To anticipate a case that will figure in Section 3, imagine that you have a cousin who is incredibly on the ball. He always shows up on time, never misses an appointment, has a calendar with every possible reminder on it, and generally is perceived as the most conscientious person in your social network. One day, though, in a very rare exception to his usual pattern, your cousin forgets to return your jacket, as he promised he would do. Further stipulate that on whatever account of the real self you favor, your cousin is alienated from the forgetting. Finally, assume that you do not really need the jacket for anything pressing. I submit that one common (which is not to say universal) reaction to this is to excuse the cousin from responsibility. We say: "Don't worry about it." If he says, with evident pangs of guilt, "I feel terribly responsible," you might reassure him that he is not blameworthy. If he continued to protest, "Well, maybe it doesn't matter much to you, but I'm still to blame," one sensible response would be, "This isn't who you really are; don't sweat it." In other words, one response would be to say that because the forgetting does not represent his real self, he is not responsible for it. Addiction and compulsion play no role in this response.

Of course, some will not want to accept a reduction in responsibility in some or all cases of alienation.[7] But as with the intuitions we have about implicit bias, exculpating reactions to alienation are sufficiently widespread and compelling that I am going to take these judgments as data, granting that they are Real Self Theory's strong suits. I want to investigate what kind of theory can accommodate them, rather than position them as controversial stances to be justified. In making this the dialectical starting point, it is notable that while Real Self Theory has been attacked from multiple directions, arguably the dominant line of criticism is that alignment with one's real self is not sufficient to guarantee responsibility, and that in addition, responsibility requires something like the agent having some sort of choice or control over her actions and attitudes.[8] The intuitions being taken as data here are neutral with respect to that dispute. Instead, all that is being claimed is that certain cases suggest that alienation can be sufficient for exculpation, that a suitable degree of disassociation between the attitude or action and the agent's

[7] For example, R. Jay Wallace (1994, 2006) thinks that identification and estrangement are best construed as conditions that affect autonomy, not responsibility. Cf. Fischer (2012).

[8] For some criticism of Real Self Theory on the control front, see Levy (2005); Wallace (1994); Wolf (1990). For the view that voluntary control in the ordinary sense is not necessary for responsibility, see Adams (1985); Hieronymi (2008); Scanlon (2008: ch. 4); Sher (2006); Smith (2008); Smith (2011). For an overview of the problem that even an ideally self-controlled person can have her *self* that does the controlling (i.e. her values, principles, and so on) manipulated by another agent, see Mele (2002).

real self can significantly diminish the agent's responsibility. So, regardless of whether representation of one's real self is sufficient for responsibility (or whether something like control is *also* required), the more minimal principle suggested by the intuitions granted here is that real-self-representation is necessary for full responsibility or, put otherwise, that alienation is sufficient for some exculpation.

That said, despite being less controversial, even this minimal principle appears to be problematic, as another line of criticism shows. Consider, first, a class of cases that Kasper Lippert-Rasmussen (2003: 371–3) calls "whim" cases. In these cases the agent's real self is neither associated with nor disassociated from nor ambivalent (or divided) towards the motivation for the act. She simply has no real attitude towards it. Nonetheless, she still might be responsible for it. If I act on a whim—I buy a lottery ticket or steal a piece of candy on an impulse—I am surely still responsible for my action, even though the motivation does not stem from my real self.

Second, and more relevant for our purposes, consider a man whose wife catches him in the act of cheating with another woman. He sensibly, sincerely, and truly says the exact same things that are said in the cases where Real Self Theory seems to get intuitive answers—cases like the unwilling thief or drug user: he did not want to do it, he loves his wife very much, he finds the whole idea of being with the other woman to be terrible. It is not something he has ever done before or will do again. It is not something he secretly takes joy in—he is disgusted by his behavior, and even gets physically ill when he thinks about it. (In this way, this cheating husband is different from Watson's (1987: 150) "perverse cases" where one loves doing what one wants even though doing so violates one's values, or Arpaly and Schroeder's (1999: 176–7) case of the person who secretly likes stealing.) He is not even sure why he did it: when he entertains thoughts about it being a reaction to a big argument at home or a slip after being lonely on the road for months, those explanations ring hollow, failing to resonate with him as expressive of his true desires, values, principles, unknown or unbidden wants, or any other element of his psyche, expressed or latent. Not only do they not strike him as such, he is self-aware—years of quality therapy have helped him reach these conclusions—and so he is *right*. It is not (only) that this man thinks that he did something wrong, that he wishes he did not do it, or that he disapproves of it; it is (also) that the act is not in any way expressive of his real self, *on whatever criteria you might choose to set the boundaries of the real self.* (That is, unless the real self is just that which produces *everything* that we do or intentionally do. But then Real Self Theory would make us responsible for everything we have ever (intentionally) done—a proposal so incredible that perhaps nobody since Hobbes has believed it. And it would make Real Self

Theory no longer able to say that the willing addict is responsible, while the unwilling addict is not.)

What is the intuitive judgment in this case? We might feel more sympathy for this alienated cheater than we would for a willing, spiteful, or ambivalent cheater. And of course there is a certain morally dubious character to the willing cheater's action that is not present in the unwilling cheater's action. Nonetheless, I submit that a very common reaction is that the husband's *responsibility* is not significantly diminished by the fact that he is alienated from both the act of cheating and whatever mysterious element of his psychology moved him to do it.[9] I would rather be the partner of an unwilling cheater than a willing cheater, and I would find it much easier to forgive the unwilling infidelity, but I certainly would not judge the unwilling cheater less *responsible* for the act of cheating. And my suspicion, confirmed (though of course hardly proven) by a handful of anecdotal data, is that this attribution of responsibility is common enough. This judgment—that this case is one of alienation without exculpation—presents a problem for Real Self Theory's minimal principle.

Friends of Real Self Theory might resist the husband's story: perhaps he is lying about his alleged alienation, or maybe he is blind to his own motivations—he is lying to himself. In that case, while it might appear that he is alienated from his action (and motivation), he really is not. But there are two problems with this interpretation of the case.

First, as a dialectical matter, it does not help Real Self Theory. The husband's alienation is a stipulated feature of the case, and for that reason we have no more reason to question the husband's alienation than the drug user or thief's alienation. Of course, alienation is built into the drug use and thievery cases precisely because it is the very thing that Real Self Theory is supposed to nicely explain. So we cannot pursue this strategy of denying the cheater's alienation without also folding Real Self Theory's strong suit. With friends like that, who needs enemies?[10]

[9] See Arpaly (2003: 130–1, 139) for a similar judgment about a range of cases where the agent is alienated from the action and the desire that prompts it.

[10] For those who want to limit the discussion to cases of addiction, which I have suggested is a mistake, consider two points. First, we can characterize the cheating husband as someone who cheats out of a sex addiction from which he is as alienated as the heroin addict and kleptomaniac are alienated from their addictions. (And for skeptics about sex addiction, imagine that there was such a thing.) Yet many still will not get the same pattern of responsibility judgments that they get with the unwilling and willing heroin addicts: the unwilling cheater might be less bad than the wholehearted one, but the unwilling sex addict is not exculpated to the degree that the unwilling drug or thievery addict is. And that is what Real Self Theory needs us to say: alienation exculpates, full stop. The problem is that what accounts for that difference in exculpation is not accounted for by a difference in alienation. Second, the proposed limitation seems to render exculpation a matter of addiction, not alienation. It is a rejection of Real Self Theory, not a preservation of it.

Second, as a substantive matter, there is a reason why we do not question these characters' alienation. While many claims of alienation are surely just misunderstandings and denials, true alienation without exculpation is a familiar phenomenon. As it happens, I consider our case in question an actuality: a real couple who shall remain anonymous here strike me as having lived through a real-life case of alienated infidelity, on every respectable theory of the real self that I know of or can imagine. But more generally, far from being merely hypothetical, stories of alienated actions and attitudes for which we judge ourselves responsible can resonate as perfectly ordinary. Alienation can even *heighten* our sense of responsibility at times. Think of the common experience of finding it inexcusable to have said something that you do not really mean, particularly when what you said hurt somebody undeservedly, just, let us say, because you are in a bad mood.[11] As the words come out of your mouth, you recognize correctly that they are terrible and do not represent your real stance, on any theory of what constitutes the real self. And nobody, neither Frankfurt nor Watson, nor your mother, nor your best friend, not even your Freudian psychotherapist, would disagree with the claim that you are alienated from your hurtful comment. Nonetheless, you deem yourself blameworthy for it. So in addition to the problem that denying the husband's alienation is tantamount to denying the same feature of the unwilling addict, thereby giving up Real Self Theory's strong suit, this interpretation misconstrues the bigger issue at hand. The problem is not that alienation is possible only in one set of cases (drugs, stealing) and not the other (infidelity, saying something we do not mean). It is that alienation resonates with us as a responsibility diminisher in one set but not the other.

Return to the case of unwilling sexism, racism, and other discriminatory attitudes. Imagine (hopefully realistically!) that you are a committed egalitarian. Imagine (perhaps realistically!) that you are also a member of an oppressed category—a woman, a Latino, someone with a disability. We can even add that you put your money where your mouth is: you are an engaged activist, making it your life's work to end the oppression your group faces. Now imagine that you discover, to your great horror, that you harbor an implicit bias against the very group whose cause you have so passionately taken up. Your implicit attitudes are revolting against your true commitments. What is the intuitive way to judge your inegalitarian biases?

If you told me that story about yourself, my first instinct would be to say something compassionate that verges on exculpation: "That's unfortunate, but don't be too hard on yourself—just look at your true beliefs and hard work done

[11] For more on implicit bias and moods, see Madva (ms.).

to advance equality. That's the *real* you." But this is a characterization of what I would *say*, a characterization of whether and how I would *hold* you responsible, whether I would *perform the act of blaming* you. It is not my judgment as to whether your attitudes *are* blameworthy, whether you *are* responsible. We sometimes conceal the latter judgments to be less harsh to our interlocutors, for a variety of immaterial reasons. So, to take the harshness confound out of the equation, re-aim the case at me, the judge. If the story were about me, I would tell myself something different, a more complete truth.

While still recognizing the true commitments of the "real me," I would also judge myself blameworthy. I would *not* question, as we did above with the cheater, whether I was truly alienated from my implicit bias. Life's dumbfounding revelations notwithstanding, I know myself well enough to know that I wholeheartedly disavow such nonsense! I know that I am alienated from it on every theory of the real self. Unfortunately, though, piles of data show that alienation is not enough for eradication. And I would have the intuition that in this case, alienation is not enough for exculpation either. To return to the initial question, note that this reaction holds for others as well as myself: while I might be harsher from the first-person perspective when it comes to the act of *holding* responsible, my judgments as to *being* responsible are the same regardless of whether the person being judged is myself or another. We are both equally responsible, despite being alienated from our biases.[12]

It might seem tempting to try to solve this puzzle by articulating a version of Real Self Theory that can accommodate these cases. But the problem at hand is not, for example, that some versions of Real Self Theory have a flawed notion of the real self. It is that sometimes alienation from one's real self—however you care to unpack that relation—appears exculpatory, and sometimes it does not. That asymmetry is incompatible with Real Self Theory's minimal principle that alienation is sufficient for exculpation. Since the problem is structural, no tinkering with the details will fix it.

In particular, it will not do to say that our implicit biases reflect some hidden aspects of our real selves, when the real self is characterized properly. For any given Real Self Theory, if it is going to use cases like the drug use and petty theft

[12] Levy (2011: 257) and Holly M. Smith (2011: 140–1) maintain that we are only responsible for attitudes that are part of our *general* evaluative stances, not isolated "outlier" attitudes. Smith's argument is, simply, that deviant implicit biases do not reflect our full psychology. But that is just to deny, without further support, the data that we do often judge ourselves responsible for deviant alienated attitudes and actions. Levy's argument is that we cannot be responsible for outlier attitudes because we hold *agents* rather than *attitudes* responsible. But this argument contains a gap: granting that we hold agents (not attitudes) responsible, it remains open what they are responsible for.

cases to claim that it can best explain why the unwilling agent is not responsible while the willing counterpart is, it will need to offer a theory of the real self to explain this asymmetry. In this way, the "real self," while it has some intuitive resonance, is a theoretical construct set out to define the boundaries of the responsibility-bearing agent. But this will be where the problem arises: *whatever* your preferred theory of the real self, if it means that the unwilling addict is alienated from her drug use (as it must, for Real Self Theory to claim this data point as a strong suit), then we can design a case of unwilling bias that will, according to that preferred theory, also mean that the agent is alienated from her bias. And yet, in contrast to the stipulated intuitions about the unwilling drug user, many will have the intuition that the agent is responsible for her unwilling bias, that the unwilling cheater is responsible for his infidelity, and that you are responsible for unwillingly saying something when in a bad mood that you do not really mean.

It might seem at this point that Real Self Theory is just false. (This is the conclusion that Lippert-Rasmussen draws from the whim cases.) But this is not a very satisfying solution either, because it is not unreasonable to think that alienation *is* what diminishes responsibility in certain cases. Alienation, after all, is the only feature that distinguishes the desperately unwilling drug user from her willing counterpart. What this suggests—again, if we continue to take these intuitions as givens—is not that the Real Self Theory is worthless, but that it is not the whole story. So rather than rejecting Real Self Theory outright, we need to figure out how to preserve what seems most intuitive in it. Our puzzle is how to explain the intuitions that alienation exculpates in some cases (drug use, petty theft) but not others (bias, infidelity, saying something you do not really mean). The only possible solution, given our parameters, is another piece of theory. We need a norm governing responsibility attributions to explain why alienation excuses the petty theft or drug use but not the infidelity or bias.[13]

[13] One might hypothesize that the issue here is that the agents we are considering have unequally developed their capacities for self-control *prior to* their transgressions. For an argument against this suspicion, see Arpaly (2003: 140–2), who focuses on ways in which self-control can be unavailable (cf. Sher, 2006). I would add to her point that in many such cases this suspicion also distorts the object of blame: we might blame the cheater for not developing a more robust capacity for self-control, but in the first instance we are simply blaming him for his cheating, regardless of whether he has or has not developed a robust capacity for self-control. Relatedly, and decisively, we can stipulate that the agents in question gave exactly the same effort to develop self-control prior to the transgressions, rendering this hypothesis unable to explain the asymmetry between them.

2 A Theory to Accommodate the Data

Nomy Arpaly (2003) holds that we are responsible (blameworthy, praiseworthy) for our actions and attitudes just when they express or reveal our good will or ill will, or at least lack of moral concern. This view might seem to hold special promise of explaining racist attitudes, since it dovetails with Jorge Garcia's prominent theory that racism is primarily a matter of expressing ill will (Garcia 1997a, 1997b, 1999, 2001, 2004). But this view fails to accommodate certain pivotal judgments. First and foremost, it will not get us the result that we are responsible for our implicit racial or gender associations, because it seems clear that such associations need not express ill will (cf. Levy 2011: 256). Arpaly (2003: 155) claims that unwillingness, as in the case of the unwilling kleptomaniac, indicates a lack of ill will, but then that is also true in the case of the unwilling racist, who is *not* excused from responsibility by simply lacking ill will. Second, even when limiting our discussion to *explicit* attitudes and actions, the ill will approach cannot explain so-called "benevolent" sexism or racism, where the agent consciously thinks that it is truly in the interests of the woman that she not be allowed to work or that it is truly not in the interests of the non-white person to make decisions for herself (Glasgow 2009; Mills 2003; Shelby 2002).[14]

At another point, Arpaly (2003: 101–11) suggests that characters like the benevolent sexist and racist either are excused because they innocently have false factual beliefs about the objects of their disrespect (that they have inferior rational capacities, for example), or are not excused because they retain their false beliefs in the light of countervailing evidence, which reveals a lack of moral concern. But this discounts some ways of being a flawed human being. We are capable of holding onto alienated attitudes that persist against our evaluative commitments and factual beliefs, rotting away against every effort to root them out and epoxy the scars. This seems particularly likely with implicit biases. To find them inside of us, we need harbor neither ill will nor false beliefs.

A more accommodating theory that I want to pursue here posits that the content of one's attitude or action can help determine whether one is responsible for that attitude or action, because it can help determine whether alienation exculpates. Call this view *Content-Sensitive Variantism*, or CSV. On CSV, the unwilling drug user's alienation excuses him *because* his alienated desire concerns drug use, while the unwilling racist's alienation is not exculpatory *because* her alienated attitude is racist. The spouse does not get off the hook, precisely because

[14] See King (2009) for the argument that quite ordinary and common cases of negligence also involve no ill will. He concludes that, surprisingly, we are not responsible for negligence. The theory I articulate in this and the next section avoids this compromise of intuition.

what he is doing is cheating on his partner. In short, CSV says that whether alienation significantly diminishes your responsibility depends on what action you performed (or omitted) or what attitude you held (or omitted).

CSV is not a comprehensive theory of responsibility, and so it does not exclude other responsibility-affecting features, like being involuntarily hypnotized. If those other features cancel responsibility all the time, then CSV can be accepted with that constraint. Still, CSV says that *alienation* exculpates only in cases with certain content. Because this is its hallmark, CSV falls under the general umbrella of variantism: criteria for responsibility apply or fail to apply in a manner that varies based not only on the conditions of agency but also on extra-agential features (Doris, Knobe, and Woolfolk, 2007; Knobe and Doris, 2010). CSV's specific claim is that one extra-agential feature that determines whether you are responsible is the content of your action or attitude—the act performed or the attitude held.[15]

CSV is a weird view, so I want to emphasize the positive motivation for it before proceeding. The basic idea is that our attributions of culpability shift in cases like those considered above: thievery and drug use on the one hand, and on the other hand, infidelity, bias, and saying something you do not really mean. Since all of our agents are, by hypothesis, alienated from the attitudes and actions in question, alienation cannot explain the asymmetries in their responsibility. And yet at the same time, alienation *is* what excuses the unwilling thievery and drug use as contrasted with willing thievery and drug use. The proffered best explanation for these seemingly inconsistent data is that the portion of responsibility that is tied to the exculpatory power of alienation is content-sensitive. The reason why alienation excuses some attitudes and actions but not others has something to do with the content of the attitudes and actions in question.[16]

[15] As Doris, Knobe, and Woolfolk point out, it is possible to convert any variantist theory into a non-variantist theory by simply sticking together all the different criteria for responsibility in one (perhaps long) disjunction. So we can say that a variantist theory holds that there is no single non-disjunctive theory of responsibility that covers all actions and attitudes. This characterization of variantism can ward off certain challenges to the view. For instance, Dana Nelkin (2007) argues that some data that drive us towards variantism can be accommodated by the following "invariantist" theory: if you act for good reasons, then you are responsible; alternatively, if you do not act for good reasons, then in order to be responsible you must have the ability to do otherwise. Even if this view can accommodate the data, it is not an invariantist position as characterized here, because the conditions of responsibility vary according to whether you act for good reasons.

[16] One other recent content-based account of responsibility comes from Scanlon (2002: 174), who writes that, in determining whether someone is responsible for an attitude, "What matters is the content of the attitudes, not their origin or susceptibility to rational control." If we construe this as a claim about what matters *to alienation's ability to exculpate,* CSV's general architecture dovetails with Scanlon's view to that extent. That alignment notwithstanding, in the next section I will explain how CSV departs from Scanlon's position. For a content-sensitive account of autonomy, though not

3 Wolf's Burden

But we need to do more work. It is not enough to say that alienation functions as exculpatory only for certain attitudes and actions and not others. We must also explain *why these* attitudes and actions, but not *those*, are the ones where alienation is exculpatory. Without such an explanation, CSV might look like ad hoc gimmickry at best, and at worse like a reductio of the very idea of responsibility. Susan Wolf (1990: 45) identifies the burden thusly: "A satisfactory theory of (deep) responsibility must not only be able to identify which agents are responsible, and for what—it must be able to explain why they are responsible, and, ultimately, why the idea of responsibility makes any sense at all." Call this *Wolf's burden*. So far, it looks like CSV fails to shoulder Wolf's burden.

The explanation that I think is most promising is that alienation is an excuse in some cases but not in others because the one set of cases bears on certain values, while the other set does not. Let us say that the values at stake in exculpatory cases are *negotiable* values, while the values at stake in the non-exculpatory cases are *non-negotiable* values. When the partner commits an act of infidelity, he has violated a non-negotiable value. Racist attitudes, even unconscious ones, violate a non-negotiable value (at least in the actual societies we know—I will note a possible exceptional hypothetical society shortly). By contrast, when the kleptomaniac cannot resist swiping the candy bar or the user seeks out one more fix, they violate negotiable values—values whose violation we are willing to compromise on in our attributions of responsibility.

So far, of course, this is just a stipulation that there are negotiable and non-negotiable values. To truly deal with Wolf's burden, we need more: we need a theory of what makes a value non-negotiable.

One promising theory is that our standards of responsibility are somehow *relational*. On Scanlon's formulation, "to claim that a person is *blameworthy* for an action is to claim that the action shows something about the agent's attitudes toward others that impairs the relations that others can have with him or her" (Scanlon, 2008: 128; cf. Scanlon, 2002: 172–4).[17] Perhaps to apply this kind of theory to some of the cases we are considering, we would have to say that the blameworthy attitude might *reveal* an already impaired relationship, in addition

responsibility (or blameworthiness), see Buss (1994). For replies to objections to variantism about responsibility in general, see Doris et al. (2007).

[17] Scanlon's overall theory has another core component that *blaming* a person (as opposed to merely judging her blameworthy) is to not only judge her blameworthy but also to take the relationship to be modified in some way as a response to the judgment of impaired relations. What I say about blameworthiness and responsibility is intended to be neutral on this other question.

to the possibility of impairing the relationship going forward, as the quoted formulation suggests.[18] But if we go with that broader formulation, which is in the spirit of Scanlon's remarks generally, then we can explain a lot of what we have been struggling to explain here.

Start with the most direct application: the case of infidelity. Monogamy is a linchpin of conventional serious romantic commitments (allowing, of course, that some couples choose non-traditional commitments). Fidelity to the commitment of monogamy represents, in many ways, commitment to the relationship itself, at least when that commitment is governed by conventional norms about how to express the commitment in behavior. For this reason, it is quite sensible that many who have been on the wrong side of infidelity find it literally unforgivable. Even those who *want* to forgive can sometimes have a hard time moving past the violation, as it represents a betrayal that destabilizes the entire relationship. And to the extent that it is possible, forgiveness, for them, involves *rebuilding* and *reconstituting* the relationship, rather than "going back to the way things were."

Bias can also be relationally corrupt. Any unwarranted biases that are systematic—whether they target race, gender, ability, height, attractiveness, or whatever—cause systemic and unjust inequality. In addition to being directly harmful, these inequalities create, perpetuate, and exacerbate social and civic rifts, the repair of which requires not just individual behavioral and institutional change but also relational reconciliation. Especially on views of, say, racism that analyze racial biases as instances of disrespect (Glasgow, 2009), biases—be they expressed *or unexpressed*—represent an affront to our status as moral co-equals regardless of their independent contributions to the perpetuation of those systems. (Perhaps in fantasy societies where such biases are not systematically implicated in these ways, the biases do not necessarily corrupt relationships, and perhaps in such fantasy worlds it would make more sense for alienation to exculpate for such biases.)

So the suggestion on the table is that some violations are inexcusable because of how they bear on our relationships. It only boosts support for this suggestion that the imputation of moral responsibility can help facilitate relational reconciliation: when bad attitudes impair relationships, they can be repaired in part by taking responsibility (Adams, 1985: 16). To return to our cases, the ability to attribute responsibility is part of moving on from the violations that happen

[18] Moreover, to be blamed, the relationship and impairment process must have a certain *valuation*. We would not blame a member of the Hitler Youth for betraying his relationship with the Nazi Party by trying to sabotage a Nazi military installment. And we would not blame a couple for putting the final nail in the coffin of a bad relationship by officially separating once they realized the relationship was already irretrievably impaired.

when cheating disrupts the very constitution of intimate (non-open) relationships and when biases disrespect our moral equality.

But do the addict and kleptomaniac not corrupt relationships too? After all, drug use can destroy families, and stealing represents a challenge to our civic equality—a disrespect of property rights that requires its own kind of repair in the justice system. In that case, the relational theory would not be able to explain the asymmetry, as this theory requires that relational corruption excludes alienation-based exculpation.

The relational theory's reply must be that our stories are underdescribed, and that addiction and theft do not create the same kind of relational crises that we find with (typical cases of) cheating or bias. It may seem that every act of stealing constitutes an affront to the relationship between victim and offender as civic equals. But it is implausible that a relationship is violated by every crime. If a poor kid steals a guitar from me, I do not see it as expressive of his sense of civic superiority over me. It is just a poor kid going through a period of desperation, and to repair the damage done I do not need my relationship with him repaired. I just want my guitar back.

What about those—more robustly described—cases where the relationship *is* impaired, say, because the theft was an expression of contempt for the victim or was part of a system of oppression that stems from bias? Here again the relational version of CSV is promising, for it will not say that alienation exculpates in such cases—relational damage has been done—which seems to be the intuitive result. Similarly, alienation will not be exculpatory for drug use exactly to the extent that it represents damage to some relationship—say, between a recovering addict and her recovery sponsor or a family that has made sobriety a condition of the commitment to one another. For those relationships, not using drugs can be non-negotiable, to the point that alienation no longer resonates as an excuse. Just like the response to infidelity, to the extent that the relationship can be salvaged after such a violation, it arguably makes more sense to see it as one that is being reconstituted rather than returned to its former state. So the relational theory can accommodate both versions of the drug use and thievery cases: in relationally corrupt versions of the story, alienation does not exculpate, but in the original relationally neutral descriptions, alienation does exculpate.

This brief tour through our featured cases shows that the relational account of CSV appears to have considerable explanatory power: it can account for the asymmetry between the (original) drug user and the thief on the one hand, and the biased person and the cheater on the other. It also can easily account for the attribution of responsibility for saying something you do not really mean, since this too damages a relationship.

Below, I will try to preserve these strengths of the relational account. But we cannot simply accept the relational model, for judgments about other alienated actions and attitudes are not compatible with it. In particular, cases of ordinary neglect, lapse, and forgetting can be brought to bear against it.

Consider the following instance. Your extended family is planning a once-in-a-lifetime, collective, international vacation. It has been in the works for a year: reservations have been made, tickets and swimwear have been bought, leave from work has been secured. The night before departure, the most dutiful member of the family—your uber-responsible cousin discussed earlier—goes to check in for his flight on-line. Recall that this person is always on the ball. He never misses an appointment, never makes a commitment he cannot keep, has a calendar with every possible date and deadline on it. When it comes to life's cancellations, failures, and missed opportunities, other people let *him* down; he does not let *them* down. He is hit like a ton of bricks, then, when he discovers that his seldom-used passport expired a few weeks earlier. He and his nuclear family will not be making the trip on time. They will be delayed for three out of the six nights, leaving themselves and the rest of the extended family significantly disappointed.

Neil Levy (2005: 13) holds that when an agent apologizes in such cases, it is a roundabout way of saying that he is not truly responsible, that his lapse is not really attributable to him. But in this case, I submit, many have another reaction: the agent *is* responsible for this forgetting, even though he is *not* identified with the act, even though it is not attributable to his real self. It goes totally against his character, behavioral record, desires, values, and anything else you want to say constitutes the real self. Moreover, it cannot be traced to some earlier culpable shortcoming: he just plain forgot, as humans do. Paradoxically, though, this alienation only seems to *heighten* his sense of responsibility for it. It certainly does not diminish it. (This diagnosis is not mere armchair speculation either, for the story is a real case. Although anonymization has been introduced to protect the guilty, what is important is that the agent in question said, unequivocally: "Of course I am responsible—I feel terrible!" and when I asked added, "Of course I don't identify with forgetting to renew my passport—it goes totally against who I am!" These two judgments—alienation and responsibility—were endorsed by all concerned parties.)

The relational account of what makes a value non-negotiable cannot handle the passport case. The traveler's family members, disappointed as they may be, are an ordinary, charitable bunch, and their attitudes towards and relationships with the guilty party do not change one bit. Their trip has been compromised, but their bonds with one another are fully intact; and though they are upset, they recognize that the agent's lamentable forgetting does not reveal or cause anything untoward in their relationships. And if the relationship has not been degraded by the lapse and the lapse reveals no prior degradation—it is just a straightforward lapse, with no relational significance—then the relational account cannot explain the judgment of responsibility.[19]

What else could explain it? I suspect that our best hope is to focus on the fact that the traveler's forgetting substantially worsened an event that held great significance for his family. That would also explain why that same agent, your cousin, is *not* judged responsible for simply forgetting to return your jacket: it does so little harm that alienation can exculpate. Or, rather, he would not be judged responsible *unless* you were really counting on that jacket for some reason; then, because of that increased quotient of harm, responsibility appears to kick in again.

Notice, importantly, that it is not always fitting to say merely that the low stakes allow us to not *hold* the agent responsible for inconsequentially forgetting to return the jacket even though he *is* "technically" responsible. I predict that one common reaction, at least, is to say things reflect the judgment that he simply *is* not "technically" responsible; the stakes are so low in this case that alienation is sufficient for exculpation. We might say, for example, "I wouldn't blame you for something so insignificant!" Note, furthermore, that the low stakes are not *themselves* sufficient to do the exculpating. It is the low stakes in combination with the alienation: if the person forgetting to return the low-stakes jacket was not alienated, and instead proudly declared to his cousin that he does not care about returning other people's property because he wants to undermine

[19] Although cases of forgetting, lapse, and negligence are used by some, such as Smith (2005) and Sher (2006, 2009), to motivate new and improved Real Self Theories, the passport case's structure means that no improvements can help Real Self Theory accommodate it. For example, this kind of forgetting cannot be handled by Sher's (2009: ch. 8) theory that the agent is responsible for forgettings because they stem from the same psychophysical structure that gives rise to our evaluative judgments, regardless of whether failures of attention like this are judgment-based. Though Sher's theory can accommodate some cases of lapse, this psychophysical structure need not be the structure that is responsible for the traveler's lapse, or, perhaps more obviously, for implicit biases that are radically detached from our explicit evaluative commitments. This can be made vivid by supposing, for argument's sake, both that the mental (psychophysical) module responsible for evaluative judgments evolved under survival and reproductive pressure to cooperate, and that a separate unconscious group-bias module evolved independently as a general in-group/-out-group mechanism whose reactions assisted in intraspecies competition for scarce resources. If this were the right evolutionary story, it would sever any ties between the psychophysical structure responsible for our biases and the psychophysical structure responsible for our evaluative judgments, but that would not change our attributions of responsibility. Of course, I do not assert that this evolutionary tale is true. I raise it only to highlight that attitudes and attitude-absences can be due to psychophysical structures that are wholly separated from the structures that give rise to evaluative judgments, or anything else with which we might want to identify the real self; yet such separation does not modify our attributions of responsibility.

modernity's corrupting materialism (i.e. he identified with his forgetting), he would be judged responsible for the lapse, despite the low stakes.

Although I put forward these anecdata and armchair predictions on their own, it is worth noting that they harmonize with experimental results showing that we are more likely to attribute responsibility for an accident the more harmful it is (Walster, 1966).[20] Additionally, this kind of harm-based CSV can also shed light on an underdiscussed dimension of alienated drug addiction. It is one thing to look past your painfully unwilling addict sister's irrepressible desires and repeated relapses. Many do think that responsibility is diminished in such cases of drug use. But it is another thing when her alienated addiction drives her to steal your retirement savings and severely neglect the basic needs of her children (your nieces and nephews!), leading to their devastating malnourishment, impaired development, and in one case death due to exposure. For many of us, such behavior is unforgiveable, even when it is alienated. It violates values that are non-negotiable. Plausibly, that is because the *harm* has pushed the behavior into the realm of inexcusability.[21]

So while the relational interpretation of CSV seems to do a very nice job explaining certain cases, the harm interpretation seems to do a better job with cases that have no relational impact, such as the passport case and its asymmetry vis-à-vis cases that differ only in harm done (like the jacket case). If there is no better explanation for the kinds of cases examined here, then we appear to have a relatively messy, but explanatorily adequate, story to tell about alienation and responsibility. On this story, alienation can diminish your responsibility for an action or attitude, but whether it does so in any given case depends on whether

[20] Importantly, Walster also showed that this feature of our attributions of responsibility does not correlate with a judgment that the person causally responsible for the act was careless. And, when the experimenters cranked up the level of harm, respondents arrived at stricter moral standards. Note also that the harm interpretation of CSV parallels other domains where a harm threshold appears to play a role in asymmetrical judgments, such as successful murder being a worse crime than attempted murder. It also fits nicely with strict liability in the law (where we can be held legally responsible for the results of, say, sufficiently dangerous behavior whether or not we are at fault for those results). All of these points of contact should strengthen the appeal of the harm interpretation of CSV.

[21] One alternative explanation for the asymmetrical judgments that we are exploring here, suggested by an anonymous reviewer, is the theory that alienation is only possible in relatively low stakes, or weakly harmful cases. (If that were true, then alienation would always exculpate, but it would only be possible in certain low-stakes cases.) Perhaps, for example, we exercise such control when the stakes are high that alienation is impossible. The case of the unwilling addict who, to her monumental and everlasting regret, accidentally lets her child die from exposure by leaving a window open at night strikes me as a compelling case that full alienation is possible in high-stakes cases as well as low-stakes cases. I say more about real-self variantism generally in the conclusion of this chapter.

that action or attitude indicates a corrupted relationship *and* on whether that action or attitude caused sufficient harm.

I want to clean up this mess a bit by reducing the relational dimension to the harm dimension, on the plausible thought that relational impairment *constitutes* a harm. If it does, then we can explain all of our cases with one principle: the husband who is alienated from his infidelity and the egalitarian activist who is alienated from her bias are responsible for their actions and attitudes because the relational corruption entailed by those actions and attitudes (even for bias that is *unexpressed*, recall) just is sufficiently harmful; the traveler is responsible because his forgetting to renew his passport causes a sufficient amount of harm; and the alienated agent who forgets to return the inconsequential jacket is not responsible, because the lapse neither constitutes nor causes a sufficient amount of harm. Similarly, agents are not responsible for ordinary unwilling desires and actions stemming from alienated addiction and kleptomania, because they fail to impair a relationship or cause a significant amount of harm; but if the harm done by those actions crosses a certain threshold (or if alienation disappears), responsibility re-enters the picture.[22]

I now want to make the harm interpretation of CSV bolder.[23] On the bold harm interpretation of CSV, *content* is not really doing the ultimate work; rather, *harm* does it all, and content is just a proxy for harm. On this way of thinking, when we talk about responsibility for *kinds* of thought and action, kinds like *drug use*, *thievery*, *bias*, and *infidelity*, we are really using heuristics where content substitutes for harm. Assessing responsibility, predicting consequences, and even merely identifying the elements of an action or attitude are complex and time-consuming tasks. A way of cutting through the difficulty is to use simplified kind terms, such as *infidelity* or *drug use.* But the kind terms are apt only because they track harm. Infidelity, for an easy example, reflects the impairment of a relationship, and if that relationship is good, then the infidelity constitutes a harm. Arguably, then, we are just using kinds as proxies for individual acts and attitudes

[22] What should we say about cases right on the threshold? We have several possibilities. Let *n* amount of harm be the threshold. We could say that some cases, where the harm is $n + 1$ or $n - 1$, really do just cross the line into or out of alienation being able to exculpate. Or we could say that *n* is a relatively wide band with fuzzy borders where there is no determinate answer whether alienation exculpates. Finally, we could identify *n* as a limit point for degrees of responsibility, such that as you get farther away from it, the ability of alienation to exculpate grows. I am not sure which of these answers is best, though I am partial to the last interpretation. But in any case, all are consistent with the harm interpretation of CSV. Another question (which I owe to Luc Faucher) is whether the harm that determines responsibility in conditions of alienation should be the actual harm done or some harm that is possible. Again, both possibilities are consistent with the harm interpretation of CSV.

[23] I am grateful to Jules Holroyd and Erin Beeghly for pressing me on issues of content and description that recommend this bolder formulation.

that are sufficiently harmful—a *useful* heuristic, to be sure, since the individuals in these kinds are typically harmful. If this is right, then the final formulation of our general principle can omit reference to *content* and simply say that whether alienation can exculpate depends on whether the act or attitude in question is sufficiently *harmful*.[24] In that case, we could trade out the harm interpretation of content-sensitive variantism for a more streamlined harm-sensitive variantism. We could give up CSV in favor of HSV.[25]

Both HSV and the harm interpretation of CSV successfully shoulder Wolf's burden. On the harm views, the reason that alienation for certain attitudes and actions fails to exculpate is that they are harmful, in the broadest possible sense. This view explains responsibility not by appealing to a descriptive property possessed by agents—although such metaphysical criteria, such as those excusing brainwashed agents, can certainly be coupled with CSV and HSV. Rather, this aspect of responsibility is ultimately grounded in the *normative* explanation that some conditions for responsibility can be determined by harm. Although the normative approach may not be the traditionally dominant way of thinking about these things, arguably the defensive burden should fall on those who want to insist that the criteria for responsibility are wholly non-normative. After all, ultimately we are looking for standards for applying a *moral* concept, which makes it likely that normativity would be built into the very conceptual apparatus itself.[26]

Finally, note that while these theories do supply some unity to our intuitions, the picture remains disunified on two fronts. First, like a child who cleans her toys by throwing them all under the bed, the messiness of alienation's exculpatory power is cleaned up only by moving the mess to the domain of harm. Instead of having two dimensions—relational impairment and harmful effects—determine when alienation can exculpate, now only harm determines when alienation can

[24] In addition to cases of blame, which have been our featured cases, there might also be cases where whether we *praise* in cases of alienation depends on whether the case has enough *benefit*. (I am grateful to Randy Mayes for this point.) Consider the widely discussed case of Huckleberry Finn failing to return Jim to his slave-owner, despite what Huck considered to be his obligation to do so, indicating alienation (e.g. Arpaly and Schroeder, 1999). It is arguable that this alienation notwithstanding, failing to turn in Jim deserves praise because it reflects a healthy relationship and nets a significant beneficial outcome. Determining whether such a broader theory of praiseworthiness can prevail, though, would require greater consideration of a variety of cases.

[25] HSV can be divided into two types: those views that say tie alienation's exculpatory power to harm for every *token* action, and those that tie it to harm for every *type* of action. I am inclined toward the token version, but I will not defend that inclination here.

[26] For more on the normative approach to identifying criteria for responsibility, see e.g. Wallace (1994). Insofar as we greatly care about reducing harm, the harm theories also mesh nicely with the claim that responsibility is best understood as a way of making us better (Vargas, 2013).

exculpate, but those two dimensions determine harm. That may be more of a bookkeeping victory than an achievement of monism. Second, and more important, this remains a variantist theory, and in that respect it is by definition not unified. It claims that alienation sometimes diminishes responsibility and at other times does not, depending on the harm done. That never changes for either CSV or HSV.

4 Conclusion

I have considered only a small sample of cases here, so the foregoing leaves open the possibility that more features of an act or attitude, besides just relational impairment and consequent harm, can affect whether alienation diminishes responsibility. At this point, however, we can say that the compelling and widespread intuitions about the cases we have considered are accommodated by variantist harm theories better than the alternative accounts that we have examined along the way.

There is, however, one final rival account worth considering: a return to Real Self Theory. On this account, there is one, invariantist criterion for responsibility: an agent is responsible for an act or attitude just when that act or attitude expresses her real self. However, there is a twist: the *real self* is given a variantist analysis, such that different sets of conditions can determine whether the agent is alienated or whether her real self is engaged. If we can work up the right conditions, likely involving a harm component, otherwise identical agents might be alienated in cases of drug use and thievery but not alienated in cases of bias, infidelity, or forgetting. This would produce the desired results that the addict and kleptomaniac are alienated and therefore not responsible, while the cheater, forgetter, and biased person are not alienated and therefore responsible.

But although this theory would accommodate our data, it is in one important respect less compelling than our variantist theory of responsibility. To make the real self variantist in this way is just to adopt a gerrymandered set of claims to match theoretical goals, without any real underlying rationale—certainly not one that captures any ordinary idea of the real self. In fact, as argued above, it strains commonsense to think that the addict and kleptomaniac might be alienated when structurally identical cheaters and biased people are not. But if the real self is divorced from any principled rationale, we lose the motivation to use the machinery of Real Self Theory. By contrast, the kind of variantism found in CSV and HSV is motivated by adherence not to some theoretical construct, but to ordinary judgments of responsibility. So, while variantist theories of the real self seem like ad hoc theoretical constructs, a variantist theory of responsibility is a plausible analysis of the ordinary concept of responsibility.

If that is right, and if no better theory of responsibility is on offer, then we stand at a crossroads. We can go with HSV (or CSV) and thereby preserve common judgments about responsibility under conditions of alienation, or we can maintain an invariantist theory of responsibility and give up those common judgments. It seems to me that those common judgments stick hard. The harm-based variantist view might be somewhat odd, but harm does matter morally speaking, and HSV does not seem so odd that it is worth giving up those judgments for oddness alone.

Acknowledgments

I am grateful to Erin Beeghly, Michael Brownstein, Luc Faucher, Jules Holroyd, Joshua Knobe, Alex Madva, Randy Mayes, Jenny Saul, David Shoemaker, Manuel Vargas, participants at the 2012 Sheffield Workshop on the Ethical and Political Implications of Implicit Bias, and an audience at Sacramento State University, for helpful feedback on earlier versions of this work.

References

Adams, Robert M. (1985). "Involuntary sins." *The Philosophical Review* 94: 3–31.

Arpaly, Nomy (2003). *Unprincipled Virtue: An Inquiry Into Moral Agency*. Oxford: Oxford University Press.

Arpaly, Nomy, and Timothy Schroeder (1999). "Praise, blame, and the whole self." *Philosophical Studies* 93: 161–88.

Buss, Sarah (1994). "Autonomy reconsidered." In Peter A. French, Theodore E. Uehling, Jr, and Howard K. Wettstein (eds.), *Midwest Studies in Philosophy Volume XIX: Philosophical Naturalism*. Notre Dame, IN: University of Notre Dame Press: 95–121.

Cameron, C. Daryl, B. Keith Payne, and Joshua Knobe (2010). "Do theories of implicit race bias change moral judgments?" *Social Justice Research* 23: 272–89.

Dasgupta, Nilanjana (2004). "Implicit ingroup favoritism, outgroup favoritism, and their behavioral manifestations." *Social Justice Research* 17: 143–69.

Devos, Thierry, and Mahzarin R. Banaji (2005). "American = white?" *Journal of Personality and Social Psychology* 88: 447–66.

Doris, John M., Joshua Knobe, and Robert L. Woolfolk (2007). "Variantism about responsibility." *Philosophical Perspectives* 21: 183–214.

Fischer, John M. (2012). "Responsibility and autonomy: The problem of mission creep." *Philosophical Issues* 22: 165–84.

Frankfurt, Harry (1971). "Freedom of the will and the concept of a person." *Journal of Philosophy* 68: 5–20.

Garcia, J. L. A. (1997a). "Current conceptions of racism: A critical examination of some recent social philosophy." *Journal of Social Philosophy* 28(2): 5–42.

Garcia, J. L. A. (1997b). "Racism as a model for understanding sexism." In Naomi Zack (ed.), *Race/Sex: Their Sameness, Difference, and Interplay*. New York, NY: Routledge: 45–59.

Garcia, J. L. A. (1999). "Philosophical analysis and the moral concept of racism." *Philosophy and Social Criticism* 25(5): 1–32.

Garcia, J. L. A. (2001). "Racism and racial discourse." *The Philosophical Forum* 32: 125–45.

Garcia, J. L. A. (2004). "Three sites for racism: Social structurings, valuings, and vice." In Michael P. Levine and Tamas Pataki (ed.), *Racism in Mind*. Ithaca, NY: Cornell University Press: 35–55.

Glasgow, Joshua (2009). "Racism as disrespect." *Ethics* 120: 64–93.

Hall, D. and B. K. Payne (2010). "Unconscious attitudes, unconscious influence, and challenges to self-control." In Y. Trope, K. Ochsner, and R. Hassin (eds.), *Self-Control in Society, Mind, and Brain*. New York, NY: Oxford University Press.

Hieronymi, Pamela (2008). "Responsibility for believing." *Synthese* 161: 357–73.

Holroyd, Jules (2012). "Responsibility for implicit bias." *Journal of Social Philosophy* 43: 274–306.

Kelly, Daniel, and Erica Roedder (2008). "Racial cognition and the ethics of implicit bias." *Philosophy Compass* 3: 522–40.

King, Matt (2009). "The problem with negligence." *Social Theory and Practice* 35: 577–95.

Knobe, Joshua, and John M. Doris (2010). "Responsibility." In John M. Doris and the Moral Psychology Research Group (eds.), *The Moral Psychology Handbook*. Oxford: Oxford University Press: 321–54.

Levy, Neil (2005). "The good, the bad, and the blameworthy." *Journal of Ethics and Social Philosophy* 1(2): 1–15.

Levy, Neil (2011). "Expressing who we are: Moral responsibility and awareness of our reasons for action." *Analytic Philosophy* 52: 243–61.

Levy, Neil (2014). "Consciousness, implicit attitudes and moral responsibility." *Noûs* 48(1): 21–40.

Lippert-Rasmussen, Kasper (2003). "Identification and responsibility." *Ethical Theory and Moral Practice* 6: 349–76.

Machery, Edouard, Luc Faucher, and Daniel R. Kelly (2010). "On the alleged inadequacies of psychological explanations of racism." *The Monist* 93: 228–54.

Madva, Alex (ms.). "Implicit bias, moods, and responsibility."

Mele, Alfred R. (2002). "Autonomy, self-control, and weakness of will." In Robert Kane (ed.), *The Oxford Handbook of Free Will*. Oxford: Oxford University Press: 529–48.

Mills, Charles W. (2003). "'Heart' attack: A critique of Jorge Garcia's volitional conception of racism." *The Journal of Ethics* 7: 29–62.

Nelkin, Dana K. (2007). "Do we have a coherent set of intuitions about moral responsibility?" *Midwest Studies in Philosophy* 31: 243–59.

Payne, B. Keith (2006). "Weapon bias: Split-second decisions and unintended stereotyping." *Current Directions in Psychological Science* 15: 287–91.

Saul, Jennifer (2013). "Implicit bias, stereotype threat, and women in philosophy." In Fiona Jenkins and Katrina Hutchison (eds.), *Women in Philosophy: What Needs to Change?* Oxford: Oxford University Press: 39–60.

Scanlon, T. M. (1998). *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Scanlon, T. M. (2002). "Reasons and passions." In Sarah Buss and Lee Overton (eds.), *Contours of Agency: Essays on Themes from Harry Frankfurt*. Cambridge, MA: MIT Press: 165–83.

Scanlon, T. M. (2008). *Meaning, Permissibility, and Blame*. Cambridge, MA: Belknap Press.

Shelby, Tommie (2002). "Is racism in the 'heart'?" *Journal of Social Philosophy* 33: 411–20.

Sher, George (2006). "Out of control." *Ethics* 116: 285–301.

Sher, George (2009). *Who Knew? Responsibility without Awareness*. Oxford: Oxford University Press.

Smith, Angela (2005). "Responsibility for attitudes: Activity and passivity in mental life." *Ethics* 115: 236–71.

Smith, Angela (2008). "Control, responsibility, and moral assessment." *Philosophical Studies* 138: 367–92.

Smith, Holly M. (2011). "Non-tracing cases of culpable ignorance." *Criminal Law and Philosophy* 5: 115–46.

Steinpreis, Rhea E., Katie A. Anders, and Dawn Ritzke (1999). "The impact of gender on the curricula vitae of job applicants and tenure candidates: A national empirical study." *Sex Roles* 41: 509–28.

Vargas, Manuel (2013). *Building Better Beings*. Oxford: Oxford University Press.

Wallace, R. Jay (1994). *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.

Wallace, R. Jay (2006). "Caring, reflexivity, and the structure of volition." In R. Jay Wallace, *Normativity and the Will: Selected Essays on Moral Psychology and Practical Reason*. Oxford: Oxford University Press.

Walster, Elaine (1966). "Assignment of responsibility for an accident." *Journal of Personality and Social Psychology* 3: 73–9.

Watson, Gary (1975). "Free agency." *Journal of Philosophy* 72: 205–20.

Watson, Gary (1987). "Free action and free will." *Mind* 96: 154–72.

Wolf, Susan (1990). *Freedom within Reason*. New York: Oxford University Press.

Woolfolk, Robert L., John M. Doris, and John M. Darley (2006). "Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility." *Cognition* 100: 283–301.

1.3

Attributability, Accountability, and Implicit Bias

Robin Zheng

1 Introduction

When trying to assess whether people are morally responsible for implicit bias, we are likely to find ourselves pulled in opposite directions. On the one hand: we are unaware of our implicit biases, we cannot control their influence, and they often undermine our considered beliefs and judgments. So they do not look like the sort of thing we can be responsible for. On the other hand, we live at a time when many people do sincerely affirm commitments to social equality, where laws already exist to prohibit discrimination, and thus where implicit bias is one likely culprit for the persistence of deep inequalities. (Consider the evidence furnished by "CV studies," which show that identical CVs labelled with typical names from different social categories—African American or White American, male or female, and so on—receive significantly different response rates and evaluations.[1]) Indeed, implicit biases are at once the *products* and the *perpetrators* of social inequality: unjust social structures breed implicit biases, and implicit biases impede structural change. And so we cannot simply let people off the hook. In this chapter I will show how we can resolve this tension. The key is a distinction between two different concepts of moral responsibility, which I lay out and explain in Section 2. Responsibility as "attributability" depends on the notion that actions are expressions of our agency. We are morally responsible for our actions in this sense only when they reflect who we are as moral agents—that is, when they are properly *attributable* to us as manifestations of our ends, commitments, or values. Responsibility as

[1] For representative examples, see Bertrand and Mullainathan (2004) and Moss-Racusin et al. (2012).

"accountability," however, depends on the social and institutional practices governing the distribution of duties and burdens across different roles and positions within a moral community. We are morally responsible for our actions in this second sense when it is appropriate for others to enforce certain expectations and demands on those actions—in other words, when it is appropriate for others to hold us *accountable* for them. In Section 3 I consider the question of whether we are attributively responsible for actions caused by implicit bias, and in Section 4, whether we are accountable for them. I thus reconcile our conflicting intuitions by showing that we can lack attributability for implicit bias,[2] but at the same time still be *accountable* for it. What this amounts to, I further argue and defend in Section 5, is that we should refrain from what I call "appraisal-based responses" in favor of "non-appraising responses." Leaving aside ascriptions of attributability and focusing on accountability will not only lead to more effective practices for mitigating the harms of implicit biases, but will also do more justice to our moral experience and agency.

2 Two Routes to Responsibility

A number of philosophers have explored the distinction (or very closely related distinctions) between attributability and accountability. T. M. Scanlon (1998), for instance, distinguishes between what he calls "responsibility as attributability" and "substantive responsibility." Gary Watson (2004) presents a distinction between the "aretaic" or "attributability" face of responsibility on the one hand, and "accountability" on the other.[3] And a number of political philosophers have proposed distinctions between backward-looking versus forward-looking, and "metaphysical" versus "practical" or "political" models of responsibility.[4] My own interpretation, which is more Scanlonian than Watsonian, shares similarities

[2] For reasons of pure convenience, I will use "responsibility for implicit bias" to stand in for "responsibility for actions caused by implicit bias." Any references to responsibility for other aspects of implicit bias will be explicitly marked. I will also focus on actions rather than judgments or beliefs. To be sure, implicit bias does operate—perhaps primarily so—on the formation of beliefs. There is undoubtedly some sense in which we can be morally responsible for our doxastic states, but to avoid such thorny theoretical issues (see Watson, 2004, for an illuminating discussion) I will restrict my discussion to actions, some of which may essentially involve beliefs caused by implicit bias, e.g. the action of hiring one job candidate over another.

[3] See Fischer and Tognazzini (2011), who offer an extensive treatment of the conceptual distinctions between, and within, attributability and accountability. Unlike Fischer and Tognazzini or Watson, however, I do not take attributability of an action to be a necessary condition of accountability for that action. See also Oshana (1997); Shoemaker (2011); Smith (2012).

[4] See Goodin (1995), Kelly (2002), Young (2011), and Baier (1987).

with these other distinctions. However, I motivate my interpretation by grounding it in a conceptual genealogy that illustrates how these two different concepts of moral responsibility arise from two fundamentally distinct sources of philosophical concern.

The first route to responsibility—attributability—begins in metaphysics and action theory. It emerges out of the "problem of action"—that is, the problem of understanding ourselves as agents given a naturalistic picture of the world where in we, like everything else, are just segments in a tightly linked causal chain. In other words, how do we make room for the idea that we are the *origins* and *authors* of our actions, and not just the arenas in which sequences of mental events occur one after the other? One prominent tradition of solutions to the problem of action suggests that we can justifiably conceive of ourselves as agents when our actions "belong" to us in the appropriate way, that is, when those actions are *authentic* or *autonomous*. On a Kantian picture, this is because rational action is distinctively subject to self-reflective awareness, and can thus be the product of deliberative choice or endorsement on the basis of reasons. In other words, we can, when moved to act in some way, become aware of this motive within us; we can then deliberate about the reasons for or against that action, and ultimately choose the grounds on which we will perform or refrain from it. It is an agent's endorsement of the principles or motives on which she acts that makes it "her" action, and that makes her into a full-blooded agent rather than a mere cause.[5] Non-Kantians may reject the emphasis on reflective deliberation and endorsement, but maintain that we are morally responsible for actions that express aspects of our agency, character, or selves.[6] Because our agency is constituted by our ends, values, and commitments—by our "practical identities," in short[7]—our actions provide the grounds for a kind of appraisal that is not available to non-rational and non-moral agents. When and only when an action is thus *attributable* to a person is she properly subject to such assessment. While agents are responsible for many kinds of actions—clumsy

[5] See for example Frankfurt (1988), Velleman (1992), and Korsgaard (2009).

[6] See for example Smith (2005), Arpaly (2004), and Sripada (2015). Part of their disagreement with Kantians turns on the difference between what we might call "will-based" and "character-based" features of an agent that can be manifest in her actions: things like her intentions, commitments, or endorsement of reasons, on the one hand, or her ingrained traits, habits, virtues, and vices on the other. However, these differences do not matter much for my purposes; they are all ways in which we understand actions as somehow reflecting "deeply" on an agent. Thus they can all be contrasted with viewing actions in terms of accountability. So I will treat will-based and character-based features more or less interchangeably. I am grateful to Michael Brownstein for pointing this out to me.

[7] The term is Christine Korsgaard's (2009).

or graceful, intelligent or not—many of these ways of being and acting are morally relevant. So this kind of "deep" appraisal,[8] deep because it attaches to the person *qua* agent, furnishes us with a first concept of moral responsibility when it takes in morally relevant features of an agent. We are morally responsible for an action in this attributability sense when that action is expressive of our morally evaluable ends, values, attitudes, or commitments. And the responses that make up such moral appraisal, like the paradigmatic (non-consequentialist) praise and blame[9] and the reactive attitudes, are what I will call "appraisal-based responses," since they reflect assessments of a person's quality of character as a moral agent.

Attributability is typically understood to depend primarily on metaphysical or psychological facts concerning the relation between an agent and her action: about the chain of (mental) events leading up to or constituting the action, or else about the attitudes and other mental states directed toward the action. Theories of (attributive) moral responsibility thus propose that we are only morally responsible for actions if we identify with the desires that moved us to them, if they express our deep selves, if they constitute the agent as an integrated whole, or if we could have done otherwise.[10] By the same token, a number of "excusing conditions" are generally accepted as specifying when actions cannot be attributed to agents; these include behaving unknowingly, unintentionally, accidentally, under coercion, or in an altered state of mind.[11] These are conditions under which a person's behavior does not flow from her practical identity—in other words, when she is not acting fully as an agent. Some beings, like young children and non-human animals, are also subject to "exemption conditions," which indicate a global lack of well-formed characters or the capacities required to reflectively deliberate and choose ends.[12]

Contrast this with the second route to responsibility—accountability—which has its origins in moral and political philosophy. It is practical through and

[8] Cf. Wolf (1990).

[9] I do not have the space to offer a full account of blame here. However, I take it to essentially involve an assessment of the agent on the basis of an action; the action "sticks" to the agent, as it were, in a way that renders judgment of the action also a judgment of the agent.

[10] Here I am alluding to views espoused by the likes of Frankfurt (1988), Wolf (1990), Korsgaard (2009), and Levy (2014), as well as in the free-will debate.

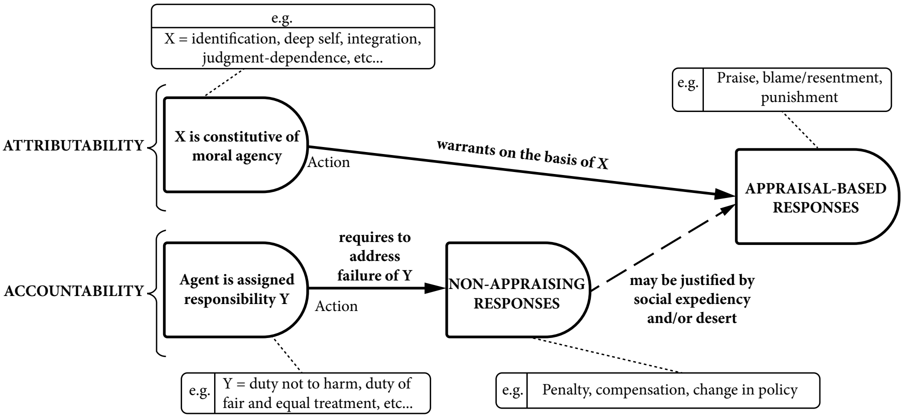
[11] I will be somewhat lax in my use of the terms "action" and "behavior," so as not to commit myself to any particular theory of action (many of which sharply distinguish the former as exemplifying certain, perhaps distinctively human, features that the latter does not.) But this in itself reflects something about the difference between attributability and accountability; for only "full-blooded" actions will in general be attributable to us, while we can be accountable even for "mere" behavior.

[12] See Strawson (1962) for the discussion of the distinction between excusing and exemption conditions.

through, and "responsibilities" show up in the first stage as the solution to a problem concerning the moral division of labor.[13] People are assigned role responsibilities for carrying out particular duties and tasks that serve our social goals (which are generally also moral goals). This notion of substantive role responsibilities thus gives rise to a second, distinct concept of moral responsibility. We are morally responsible for an action in this accountability sense when it is appropriate for others to hold us to certain expectations and demands regarding our duties and tasks—and to sanction us when we fail to carry them out. When a person's action brings about some negative consequences for others, this generates a social problem that simply cannot go unaddressed. These costs must be picked up somehow and by someone, even if there is no bad intention or fault on the part of the person involved, because there are victims who deserve redress. This means that under a fair system of distributing burdens, it will often be appropriate for the person who performed the action to bear a large share of the costs: she can be asked to compensate for damages, make reparations, or to change her practices to prevent future failures. But notice that none of this requires an assessment of character, intentions, attitudes, or values—she can justifiably be required to bear (at least some of) the costs of her behavior whether she performed them out of malice, negligence, or sheer (non-culpable) ignorance or accident. These ways of holding agents accountable for their actions are thus what I call "non-appraising responses." In a traffic accident, for instance, it is appropriate for the driver (or her insurance company!) to pick up the costs of the damage, whether she was driving recklessly or perfectly responsibly in adverse conditions; she ought to do so either way, and it is a further, separate question whether there should be sanctions on top of that. In other words, a person's behavior need not reflect anything about her at all in order for her to be appropriately required to deal with its consequences. For social and moral reasons, we might want to *further* blame or punish someone for her failure to fulfill her responsibilities. This could be justified, say, if it were expedient (i.e. according to some consequentialist), or only if it were genuinely warranted in addition to being expedient.[14] But we can also stop short of doing so.

[13] I am indebted to Elizabeth Anderson for this way of putting the point.

[14] My own view is any plausible theory which requires appraisal-based responses to be genuinely warranted will need to revert back to questions of attributability: condemnatory appraisal-based responses are only deserved if the action is genuinely expressive of or reflects on the agent herself. Strawsonian proponents of appraisal-based responses still have recourse to notions of attributability, e.g. appeals to the "quality of will," for Strawson (1962), or desert, for R. J. Wallace (2004: 107). But attributability here still plays a secondary role, relative to the primary goal of maintaining social relationships and institutions. It is only because human relationships so often depend on knowing the true intentions and characters of others that these become relevant.


e.g.
X = identification, deep self, integration, judgment-dependence, etc...
ATTRIBUTABILITY
X is constitutive of moral agency
Action
warrants on the basis of X
e.g. Praise, blame/resentment, punishment
APPRAISAL-BASED RESPONSES
ACCOUNTABILITY
Agent is assigned responsibility Y
requires to address failure of Y
Action
NON-APPRAISING RESPONSES
may be justified by social expediency and/or desert
e.g. Y = duty not to harm, duty of fair and equal treatment, etc...
e.g. Penalty, compensation, change in policy

Note that we arrive at appraisal-based responses with both kinds of responsibility, but along different paths. (To better orient myself in relation to the literature: I construe both consequentialists and Strawsonians as taking the accountability route, where moral responsibility is defined firstly in terms of the aptness of particular social relationships and practices. Non-Strawsonians, by contrast, take the attributability route: they are fundamentally and directly concerned with whether metaphysical relations hold that would make it the case that people deserve praise, blame, and punishment.[15]) Accountability is thus primarily a matter of interpersonal, not metaphysical, relations; not what it takes to be an agent, but what it takes to be a member of a community of agents.

As a last way to illuminate the contrast between attributability and accountability, consider another commonly invoked distinction between "being responsible" and "holding responsible." In my sense, roughly, attributability is about being responsible and accountability is about being held responsible. I say "roughly," however, because the difference is more a matter of priority than definition. We can *hold a person attributively responsible* by praising or blaming her, but this is only licensed when she is responsible in the right way—that is, when her action flows from her practical identity. Conversely, a person *is accountable* only when she is being held to an appropriate network of expectations, demands, and responses by others in her community. In the former case, holding responsible depends on first being attributively responsible; in the latter, being responsible depends on first being appropriately held accountable.

My aim in this section has been to clarify the two distinct families of questions that each *concept* of responsibility raises, without yet providing substantive or

[15] I am grateful to Neil Levy for pushing me to clarify this point.

comprehensive *conceptions* of attributability or accountability that would answer those questions.[16] Similarly, in what follows I will sketch out the contours of a view about responsibility for implicit bias, without grounding it in any particular first-order theory of attributability or accountability.

3 Attributability of Implicit Bias

Recall that attributive responsibility is thought to require some sort of special metaphysical or psychological relation between an agent and her actions. It is this concept of responsibility that we have in mind, I suggest, when we are inclined to let people off the hook for implicit bias. To show this, I will reconstruct what I call "The Simple Argument" against ascribing moral responsibility for implicit bias. I will not be arguing that the argument is sound, but that it is instructive to examine the motivations behind it.

The strength of the argument, as I see it, rests largely on two standard intuitions about moral responsibility. The first is this:

The Distinctiveness Intuition
We are responsible for our actions and attitudes in a way that young children, non-human animals, and other such beings are not.

I take it that this is uncontroversial: though the difference may be one of degree rather than kind, and though it is difficult to pinpoint exactly what it consists in, it seems clear that non-human animals and children are not fully agents—or at least, not fully moral agents—in the way that normal human adults are. The second intuition is as follows:

The Endorsement Intuition
There is some important moral difference between a person who would endorse the influence of an implicit bias on her judgment and behavior if she were reflectively aware of it, and some other person who would reject such an influence.

Again, this seems absolutely correct—indeed, vitally important—because to deny that there is any difference whatsoever between the two cases is to lose sight of the features that make us agents at all, or that make it possible (at least in part[17])

[16] I follow Rawls (1999) in distinguishing between a "concept," which is the general specification of some notion and the function it performs, and a "conception," which is a particular substantive theory of what realizes or instantiates that concept.

[17] Note, for example, that even Smith (2004), who explicitly rejects the view that moral responsibility requires conscious endorsement, still restricts moral responsibility to attitudes that are dependent on our judgment, that is, which rest on our rational evaluation of reasons and for which it makes sense in principle to demand justification.

to have fully developed moral characters: our capacity for reflecting on and choosing our ends. These two intuitions lead naturally to the following argument:

The Simple Argument

1. We are attributively responsible for our actions when and only when we stand in some special relation to them.
2. We do not stand in such a relation to our actions when they are caused by implicit biases that we would not endorse.
   Therefore,
3. We are not attributively responsible for actions caused by implicit bias.

The Simple Argument, simple though it may be, is actually rather compelling. Premise 1 is supported by the Distinctiveness Intuition that, since young children and non-human animals are not morally responsible for their behavior in the way we are, there must be something special about normal adult human action. The Distinctiveness Intuition also drives Premise 2: since implicit biases are one of the older automatic processes that we share with other species, and *not* the more recently-evolved deliberative processes distinctive to normal adult humans, we have reason to think that actions caused by implicit bias lack that special, distinctive feature. The Endorsement Intuition further buttresses this line of thought by directing us toward what the special feature might be: it must be somehow related to our distinctive capacity for choice and reflective endorsement of (influences on) our actions. This schema can then be adapted to a variety of candidates for the special responsibility-conferring relation.

Jennifer Saul (2013), for instance, has put forth versions of the Simple Argument in which she contends that people are not blameworthy for having (and, presumably, acting on) implicit sexist biases under the following conditions: when they are completely unaware of them, when those biases result solely from living in a sexist culture, and when mere awareness does not enable people to immediately control them. Each of these can be understood as specifying some way in which the special relation fails to hold in line with the Endorsement Intuition: we do not endorse our implicit biases if they result solely from external cultural forces, we cannot endorse them if we are not aware of them, and our inability to control them means that they operate without our endorsement. Lawrence Blum also suggests that we may not be responsible for implicit stereotypes; he states that “the entirely automatic and cognitively uninvested character of stereotypic associations” allows for “at best an extremely minimal epistemic, and moral, responsibility” (Blum, 2004: 270). Again, we should understand deliberativeness and cognitive investment as candidates for the special relation of endorsement that must hold between us and our actions if we are to be attributively responsible for them. And Neil Levy (2014) has defended the view that consciousness of reasons for action is required for moral responsibility because it functions to integrate disparate cognitive processes in such a way as to produce an agent acting upon a unified set of attitudes, through reasoning, planning, and executing projects over time; in other words, integration—which requires consciousness—is necessary for there to be a moral agent at all. Levy argues that because the non-conscious nature of implicit associations precludes them from being used in rule-based reasoning or diachronic planning and execution, the actions they cause cannot be attributed to a unified agent but only to some narrow subpart of her, and *she* is thus not morally responsible for them. To be sure, other philosophers have argued persuasively against just these sorts of proposals. But this points to another virtue of the Simple Argument, the fact that it can be filled out in many different ways. If awareness, control, non-automaticity, cognitive investment, and consciousness cannot quite do the job, it is easy to remain steadfast in the belief that the elusive special feature or features—whatever it is, or in whatever combination is required—is also missing from actions caused by implicit bias.

I will not stake out any claims about whether the Simple Argument can ultimately be rendered sound, though I think there are reasons to doubt that it can. For one thing, as more and more candidates get eliminated, the argument may begin to lose some plausibility.[18] Moreover, as Nomy Arpaly (2004: 145) has argued, it may turn out that the difference between normal human adults versus children and animals can be explained much more prosaically, without any appeal to complex notions like autonomy or authenticity as a condition of moral responsibility. Children and animals may lack the intelligence and communication skills required to engage in moral life at all—that is, they may fail to satisfy a *precondition* for moral responsibility, not just a condition of it. From the fact that children and animals do not reflectively endorse their behavior (or participate in whatever special relation is proposed), we should not be too quick to draw the conclusion that *we* are also not morally responsible for our actions when they lack reflective endorsement (Arpaly, 2004: 147).

I will thus argue for a weaker claim that still accounts for our standard intuitions. I claim only that we are not *always* attributively responsible for actions caused by implicit bias. It is worth noting at this point that I have been referring to actions “caused” by implicit bias; this is to be distinguished from actions that have merely been somehow influenced by implicit bias. To say that an action is “caused by implicit bias,” in my usage, is just to say that it *would have turned out*

[18] See Holroyd (2012) for an excellent catalogue of just such arguments.

differently were it not for the influence of the implicit bias.[19] Given the pervasiveness of implicit associations in our cognitive economies it is likely that many or most of our actions are subject to the influence of unendorsed implicit biases, but this does *not* mean we lack moral responsibility for most of our actions. So I restrict my attention to cases where the implicit bias is actually difference-making. It is in these cases (and under the two conditions proposed below) that implicit biases constitute what Frankfurt (1988: 61) calls "external" psychological forces: forces which act on us and with which we do not identify. Since we are only "passive bystanders" to actions caused in this way, they cannot be properly attributed to us. (Contrast this with a case in which an action is merely influenced by an implicit bias—that is, where the person would have behaved in the same biased manner even without the presence of that implicit bias. I take it that our intuitions in this case would be that the biased behavior can still be properly attributed to the person. This parallels the lesson of Frankfurt-style free will cases: we think a person attributively responsible for an action even if she could not have done otherwise if it is also the case that she would have acted the same way even if she *could* have done otherwise.) Actions genuinely caused by implicit bias resemble cases of hypnosis or irresistible addiction (if these exist), or sheer inadvertent accident, where we intuitively do not find people attributively responsible. While it will often be difficult to detect whether a particular action on a given occasion was genuinely caused by implicit bias, I want to emphasize that we have good reason to think that cases of outcome-shifting bias do occur—and if they do, that they matter very much. Research on aversive racism (Pearson, Dovidio, and Gaertner, 2009), for instance, has shown that people usually do not manifest racial bias in judgments where the evidence is clear and uncontroversial, and where failure to make the proper judgment could easily be chalked up to racial attitudes. But when evidence is ambiguous enough that a non-racial explanation is possible, people *do* manifest implicit racial bias in their judgments.

Without much further argument,[20] then, I propose that we are not attributively responsible for actions caused by implicit bias when both of the following

[19] Per fn. 2, an action caused by implicit bias (e.g. hiring one job candidate over another) could be one that would not have been performed if a judgment essential to it (e.g. believing that she was more qualified) would have been different were it not for the influence of an implicit bias.

[20] It bears note, however, that my proposal fits well with contemporary social psychological theories of how people actually do process information in forming judgments of blameworthiness. Malle, Guglielmo, and Monroe (2014), for instance, propose a bifurcated process by which people attempt to discern either 1) the agent's reasons for acting, if the event was brought about intentionally, or 2) whether the agent should have (i.e. obligation) and could have (i.e. capacity) prevented the event, if the event was brought about unintentionally. My first condition effectively establishes that

conditions hold. (Note again that I am not trying to offer a full theory of attributability for implicit bias here; I am only proposing the existence of *some* cases where we lack attributability, which I believe must satisfy these conditions.)

1. The agent would not upon reflection endorse the influence of the difference-making implicit bias on her action.
2. The agent has done what she can reasonably be expected to do with respect to avoiding and responding to the implicit bias.

The first condition captures the standard intuition that reflective endorsement plays some important role in moral responsibility. If an agent *would* endorse the influence of an implicit bias on her action, then that is an indication that the action *is* an expression of her practical identity.[21] Since it is properly attributable to her, she is morally responsible for her action, whether or not she can control or be aware of it.

The second condition is intended to screen out what Fischer and Ravizza (1998) call "tracing cases," in which an agent is indirectly responsible for some action because it can be traced back to other actions for which she is directly responsible. For example, a drunk driver is still attributively responsible for causing a traffic accident, but a driver who hydroplanes in pouring rain is not, *ceterus paribus*. We can view the first driver's causing the accident as a downstream effect of her agentic values, ends, etc. which produced the choice to drink so heavily, and we could justifiably go on to conclude that she was reckless, imprudent, or inconsiderate of the safety of others. On the other hand, we cannot draw such conclusions about the second driver because the car accident does not tell us anything about her values, ends, etc. at all, and because she otherwise behaved reasonably. (To anticipate my argument in the next section, note that both drivers can be expected to help pick up the costs of the accident. This is a matter of accountability, not attributability.) Analogously, then, a person who has made reasonable efforts to respond to or avoid implicit biases is not attributively responsible for such bias when it influences her action despite these efforts (though she will still be accountable).

the event was brought about unintentionally due to involvement of an unendorsed bias. My second condition invokes both an agent's obligation and capacity to prevent the influence of bias, since it is her obligation that grounds our expectation that she do so, where that expectation must be reasonable in light of her limited capacity to do so. Similar remarks apply to Alicke et al. (2008).

[21] I say that lack of endorsement *indicates* that the action is an expression of the agent's practical identity; it does not necessarily *constitute* that expression. Here again I am trying to be ecumenical with respect to different theories that could ground these conditions, by allowing that endorsement might be an evidential rather than a criterial condition on responsibility. I am thankful to Chandra Sripada for discussion on this point.

Can I say more about how to determine what a person can "reasonably be expected to do" with respect to avoiding and responding to the implicit bias? Yes, I think, but I can only hazard a few cautious claims here. While it *may* be possible to completely eliminate or avoid acquiring any morally objectionable implicit biases, in many cases this would require an unreasonable amount of time, effort, and resources. Children as young as six years old show evidence of racial implicit biases (Baron and Banaji, 2006), and cannot be expected to seek the social environments required to avoid them. It is only slightly less unreasonable to expect parents to provide these; minimizing children's exposure to gender stereotypes, for example, is notoriously difficult. The costs of, say, refusing a kindly relative's birthday present or switching out one's circle of friends might be high. On the other hand, we can be expected to stay vigilant when deciding to engage in activities or enter environments that are especially likely to foster certain implicit biases, such as consuming forms of media that are known to rely on stereotypical tropes or stock characters. And once we are made aware of our own implicit biases, we can certainly be expected to take measures that reduce their impact, like implementing anonymous evaluation procedures or exposing ourselves to counter-stereotypical conditioning.[22] (Again, though, the costs of completely eliminating their impact—say, by forgoing all face-to-face talks, conferences, and interviews—might be too high.) Preventing and responding to implicit bias is thus an imperfect duty: we cannot never try to do it, but we are not obligated to spend all our time and resources on it. As with any imperfect duty, gauging precisely when it is fulfilled is an extremely difficult matter that I cannot hope to settle here.[23] It is my view, however, that we would do well to focus our energies on elucidating this criterion of reasonable expectation for general duties of preventing and eliminating bias, rather than getting preoccupied over whether people are attributively responsible for particular instances of manifesting implicit bias. Indeed, this second criterion—with its reference to duties and expectations—forms the bridge which conjoins responsibility as attributability to responsibility as accountability, to which I turn in the next section.

[22] See, for example, Dasgupta and Greenwald (2001) and Olson and Fazio (2006). Ultimately, however, it seems to me that we would should focus more on changing the structural conditions that perpetuate implicit bias than on de-biasing individuals.

[23] Indeed, I am inclined to think that there may be a deeper kind of *moral* indeterminacy here. If there is no clear demarcation (outside of clear cases at the extremes) of when our imperfect duties are or are not fulfilled, and if it is in the nature of morality itself that there be such indeterminacy (cf. Baron 1987), then there may simply be no fact of the matter whether a particular agent is attributively responsible for some particular act caused by implicit bias. This would be different from cases where a person is blameworthy (i.e. when she genuinely does not satisfy the conditions for excuse), but we are not justified in blaming her because we lack the evidence.

4 Accountability for Implicit Bias

I have argued that there are some limited conditions under which we are not attributively responsible for implicit bias. Now I claim that, even under those conditions, we can still be *accountable* for them. We are accountable because it is appropriate for us to clean up after our own actions when a mess has been made—spilled milk has got to be wiped, though we need not impugn a person's character just for having spilled it! This need to make amends is particularly urgent when the "spilled milk" is the harm suffered by the victims of implicit bias. After all, from the victim's perspective the damage is done whether anyone is attributively responsible for it or not: harm is harm, and she is owed compensation, apology, and redress.[24] Still, we can often make such demands without invoking appraisal-based responses.

Conside Scanlon's discussion of substantive responsibility, which roughly corresponds to my concept of accountability (1998: 266). Scanlon uses an example of strict liability to show that a person can be substantively responsible for some action without being attributively responsible. In his example, a milk vendor accidentally sells some contaminated milk, through no fault or negligence on her part. But since there are laws against selling contaminated milk, it is still appropriate for her to bear the costs of having done so: she can be required to recall her product, pay a fine, compensate those who were harmed, change her inspection practices in the future, and so on. This is so even though she had no bad intentions; indeed, even though she did not act deficiently in any way. On the strict liability model, she remains accountable because she had the opportunity to avoid incurring such burdens, i.e. she could have refrained from entering the milk business. Entering the milk business, in effect, is entering what Scanlon calls an "affected area": those who do so lay down their rights not to be penalized for violating the regulations that govern the affected area. It is worth noting that because these penalties are given regardless of fault, strict liability requires strong justification; the social goals (e.g. public safety) it promotes must be sufficiently important to outweigh the risks imposed on people who enter the affected area.

Strict liability is a simple case of accountability without attributability. And it seems clear that many morally pressing situations that involve implicit bias can be read as calling for strict liability. Where people enter into positions that call for them to make decisions regarding others' merits and ability—hiring, promotions, evaluations, admissions, and the like—they accept responsibility (as accountability) for that position and lay down the right not to be penalized for failing to

[24] I am grateful for discussion with Kristie Dotson on this point.

uphold institutional standards of decision making. This line of argument is manifested in "disparate impact" clauses of anti-discrimination employment law, which operate when it can be shown that a policy differentially affects two groups that are entitled to equal consideration, and that there is some other policy that could have been implemented without leading to such different outcomes. No intent to discriminate need be shown; the policy does not have to reflect prejudicial motives or anything at all about the employer. Indeed, as I suggested earlier, these may very well be cases satisfying the conditions under which agents are not attributively responsible: they may have been caused by biases which the employer does not endorse. What matters is simply that people have been harmed in a way prohibited by law, and this demands redress. Strict liability here is justified by the great importance of the goals of fairness and social equality promoted by such anti-discrimination laws.

Strict liability when it comes to implicit bias in such domains as hiring and admissions is thus a straightforward instance of holding people accountable for implicit bias. But what about harder cases? What about more ordinary, everyday instances of harmful actions brought about by the influence of implicit biases? I have in mind here what Rowe (2008) and Sue (2010) call micro-inequities or microaggressions, prototypical examples of which include clutching purses, crossing the street, avoiding eye contact or otherwise awkward non-verbal interactions, and jokes that are not maliciously intended but still alienating. Such quotidian situations, far from being "affected areas" that one could choose to enter or not, are ubiquitous and unavoidable, and the duty to avoid harming others is not a burden one can refuse. Many of these behaviors may be subtle and unconscious, but they cumulatively have a real and large effect on the lives of historically disadvantaged groups. Thus, people on the receiving end of such treatment—and I hasten to add that the same person can be on both receiving and giving ends—are undoubtedly harmed. Yet it seems clear that these cases should not be dealt with using legal sanctions; such legislation would be ineffective at best and draconian at worst. On the other hand, if I am right that we often cannot be assured that people are attributively responsible for their microaggressions, then resentment, blame, and other appraisal-based moral sanctions are also not justified. What this points to, I think, is a paucity of more nuanced forms of moral communication. Given what we now know about non-conscious, unendorsed influences on ordinary reasoning and behavior, we should expand our moral repertoire accordingly to capture the shades and varieties of these more complicated processes. What we need, in effect, are more ways of holding people *accountable* for their biases without *attributing* those biases to them—to engage in moral criticism that does not amount to accusations of racism, sexism, or condemnations of bad character. This is too large a task for me to take on here, but I believe that developing such modes of critical moral response[25] can play an important role in making progress toward social equality. In any case, the justifiability of appraising practices of accountability for implicit bias is what explains our intuition that people should not get off the hook for actions that contribute to social inequality, even when those actions are unknowingly influenced by unendorsed biases.

5 Accountability Without Attributability

My arguments in the previous section reveal a way to escape the tension I outlined at the beginning of the chapter. To sum up, I have distinguished between two different concepts of moral responsibility, which account for our intuitions with regard to moral responsibility for implicit bias. I argued that there is a set of conditions for excuse under which we are not attributively responsible for actions caused by implicit bias, but that even then we can still be held accountable for them. This, then, is the solution: *we ought to hold people morally accountable, but not attributively responsible, for actions caused by implicit bias.* In doing so we can resolve the tension between our divergent intuitions that we both are and are not responsible for implicit bias. We can also dissolve a dilemma that Cheshire Calhoun has posed between what she calls the “justification” and “point” of moral responsibility.[26] Calhoun (1989) is concerned about “abnormal moral contexts” where certain kinds of wrongdoing are socially accepted and normalized, and moral knowledge is only available to a

[25] I borrow this term from Elise Springer. A number of philosophers have already issued a call for more nuanced moral criticism; see, for example, Smiley (1992: 254) on “a form of accountability that falls somewhere between legal punishment, on the one hand, and intrinsic moral guilt, on the other,” and Moody-Adams (1994: 303) on the stance of the “forgiving moralist.” Others have started developing such possibilities, like Fricker’s (2007: 104) “resentment of disappointment” and Springer’s (2013) model of critical moral responses as communicating moral concern. For Springer, moral criticism should not be understood merely in terms of passing verdicts or regulating behavior, but as a way of drawing attention to a moral problem for practical purposes of solving it. Another reason for developing such new modes of moral criticism is that they will facilitate the necessarily collective project of large-scale structural transformation—of not just conforming to, but also *reforming* our current distribution of duties and burdens—which is what will ultimately be required to eliminate pernicious implicit biases at their source. On my view, the task of collectively organizing to reform the system is one for which we are all accountable (cf. Young, 2011). But this means that we will need to make use of non-appraising responses that remind and motivate people to engage in this task, rather than appraisal-based responses like blame and punishment, which may be inappropriate for people who are simply conforming to the currently accepted system, or else ineffective at motivating commitment to long and difficult struggle.

[26] Cf. also Langton’s (2001) distinction between the “accuracy” and “usefulness” of resentment.

select enlightened subgroup, e.g. our current context with respect to implicit bias. Under these conditions, she thinks, we face a dilemma: a person's ignorance is a genuine excuse that frees them from blameworthiness, but on the other hand, failure to blame them defeats the point of moral responsibility altogether by allowing the problematic social norms to persist. But we can see from what I have argued that this is a false dilemma: normalized ignorance excuses from *attributability*, but not accountability. We can still engage in non-appraising responses that educate and exhort people to behave in ways that would combat structural injustice and implicit bias.

It may be objected at this point that, even if everything I have said so far is correct, I have not shown why we should refrain from ascriptions of attributive responsibility for implicit bias *in addition to* accountability, especially since I have been at pains to emphasize that excuses may exist only in a limited range of cases. This complaint may be reinforced by the thought that it would be overly sanguine to think that there do not still exist large segments of the population who *would* endorse their implicit biases, or who have not otherwise behaved reasonably with respect to avoiding and responding to them.[27] Evidence from a 1998 study by Patricia Devine and E. Ashby Plant, for instance, showed that individuals committed to egalitarianism for its own sake ("internally motivated") showed lower levels of implicit bias than those committed to egalitarianism out of concern for others' approval ("externally motivated"). Furthermore, Gawronski, Hofmann, and Wilbur (2006) argue that there is a sense in which many individuals *can* be aware of their implicit attitudes. This "content awareness" is shown by the variability in correlation between explicitly reported attitudes and implicitly measured attitudes; for example, individuals subject to a "bogus pipeline" manipulation where they are told that experimenters can detect false statements subsequently report attitudes that more closely match their implicitly measured attitudes. This again suggests that some people really would in their heart of hearts endorse their implicit biases. Given this evidence, why should we not try to sort out, for the purposes of appraisal-based responses like blame and the reactive attitudes, the people whose implicit biases really are attributable to them?

First, I want to point out that on my account, externally motivated egalitarians really are likely to count as attributively responsible for their implicit biases. To the extent that their implicit biases are linked with other non-egalitarian attitudes, it is less likely that their actions were caused by implicit bias, since these other attitudes could have produced the action even without the additional presence of the implicit bias in question. This evidence also suggests that

[27] I am grateful to Lawrence Blum for discussion of this point.

externally motivated egalitarians might actually endorse rather than reject the influence of their implicit biases, say under a different description ("non-Whites are less likely to be competent" rather than just "non-Whites can/should be discrimanted against"). But this would violate my second condition for excuse, rendering those biases properly attributable to them. And finally, even those with content awareness of their biases might ultimately still satisfy my conditions for excuse. The same paper presented evidence that people with content awareness still lack "impact awareness," which is the awareness of how an attitude influences other psychological processes. Thus a person who knows about her biases might still reject the influence of such a bias on some piece of reasoning, for instance if her intent is to be an impartial judge. (It might be said that the action still reflects an agent's flaws if the bias ends up influencing her after all, but this seems beside the point. It would be unfair to blame a person for treading on your foot by total accident and genuinely against her own volition, even if she secretly believed that you deserved to have your foot trodden on. The point, in both cases, is that the lack of intention makes it less of a full-blooded *action* genuinely attributable to the agent.) In reality, however, it seems very unlikely that anyone who endorsed the content of their bias would not fail to satisfy the second condition for excuse: that of having done what could reasonably be expected to avoid developing a bias. So for the most part, people who endorse the contents of their biases are likely to be attributively responsible for actions caused by them.[28] In the rest of this section I will present two positive arguments in defense of the claim that we ought to refrain from ascriptions of attributive responsibility. One is moral, the other pragmatic, and they work in tandem. On this view, we should refrain from appraisal-based responses for *moral* reasons if there is some chance that an agent is not attributively responsible (as is likely to be the case if she is internally motivated) but we should also refrain from appraisal-based responses for *pragmatic* reasons even if the agent is attributively responsible (e.g. if she is externally motivated). In either case, since we should always refrain from appraisal-based responses, we should not bother with attempts to determine the attributability of particular instances of implicit bias.

The moral argument is that we should err on the side of caution when it comes to engaging in appraisal-based responses. On the conditions for excuse that I have proposed, it will be nearly impossible, epistemically, to determine whether they hold in particular cases. The counterfactuals involved, which must account for the myriad influences that culminate in action, will be extremely complex, and it is unlikely that even our best neuroscience will be able to achieve the fine-grained

[28] I would like to thank Eric Swanson and Brian Weatherson for pushing me to clarify this point.

assessments needed to determine conclusively whether a particular action on a given occasion was really caused by a particular implicit bias. Of course, when we have good evidence *on other grounds* that implicit biases are attributable to a person—for example, if she demonstrates repeated offenses without repentance—then we are justified in blaming or resenting her. This is why I noted that we ought to spend more of our efforts on understanding and promoting imperfect duties of awareness, prevention, and response to implicit bias, because these can be invoked without getting entangled in complicated psychological investigation. But otherwise, it is a serious thing to engage in appraisal-based responses, especially when it involves socially sensitive domains such as race and gender, or when these judgments could be hypocritical.[29] For those who are sincerely and internally motivated egalitarians, ascribing attributive responsibility for unwanted implicit biases is too harsh. It would be like accusing a cashier of trying to cheat you on the basis of an innocent arithmetical error, or derogating a devoted but inept violinist ("Sounds like you never practiced a day in your life!"). It would be inappropriate in these cases, I take it, because such remarks constitute something like an in dignity, a kind of disrespect for a person's honest efforts. Or perhaps more aptly in the case of implicit bias, it would be like blaming a person for a behavior that they acquired as the result of some trauma, which gets triggered under certain circumstances; while such a disposition is something to be managed by her and others, it is not something for which she deserves blame or deep moral criticism.[30] (Indeed, it is likely that we are all unavoidably scarred by moral flaws: this is a tragic reflection of a deeply unjust society.) We all fall short of our ideals—and should be informed when we do—but being informed so need not amount to being condemned. Holding people accountable by demanding amends or revised behavior is enough. In fact, the people who do hold genuine egalitarian commitments ought to embrace the latter. Moreover, if such responses to discrimination were totally detached from accusations of racism, sexism, and the like, it is plausible to think that many more people would be ready to accept them and strive to do better. I believe this to be especially true of the everyday harms I discussed earlier; it remains frustratingly difficult to convey that jokes, language choice, and many other aspects of unreflective daily life can be morally objectionable even if they are not maliciously intended.

[29] Appraisal-based responses might be hypocritical if we lack the moral standing required for them, e.g. if we ourselves are (as is likely) prey to the same failings. See Hieronymi (2004) and Smith (2007) for the view that justifiable praise and blame require a certain moral standing.

[30] Here, with respect to this particular trait, we adopt the Strawsonian "objective" attitude, the attitude we take towards non-human animals, young children, and beings that are not fully moral agents. I am indebted to Sarah Buss for this example.

This last observation puts us in range of the pragmatic argument for refraining from appraisal-based responses. As Saul (2013) puts it: "If acknowledging that one is biased means declaring oneself to be one of those bad racist or sexist people, we cannot realistically expect the widespread acknowledgement that is required. Instead, we'll get defensiveness and hostility." Indeed, Czopp, Monteith, and Mark (2006) found that high-threat confrontations (where the subject was accused of sounding "like some kind of racist") induced higher levels of denial and resistance, negative affect, and negative evaluations of their confronter, as compared to low-threat confrontations (where the subject was pressed to be "a little more fair"). Appraisal-based responses thus carry the cost of engendering defensiveness and hostility, which must be overcome in order to establish and implement better practices. Although both groups of subjects in Czopp, Monteith, and Mark's (2006) study subsequently showed less stereotypical responses after the confrontation, other studies suggest that we should not conclude that high-threat responses will always be effective, particularly in the long term, and if directed at externally motivated egalitarians. Legault, Gutsell, and Inzlicht's (2011) study on the ironic effects of anti-prejudice messages showed that participants presented with a pamphlet detailing external reasons for controlling bias showed *higher* levels of both explicit and implicit bias than participants who received no intervention at all. Plant and Devine's (2001) study on internally and externally motivated egalitarians showed that, among the latter, pressure to treat Blacks favorably actually provoked a backlash: in addition to experiencing greater feelings of anger and threat, externally motivated egalitarians stated that they would be less likely to comply in the future after the removal of direct pressure. This last result is supported by research on the conditions necessary for internalizing norms. These conditions are autonomy (the need "to self-organize experience and behavior and to have activity be concordant with one's integrated sense of self"), competence (the need "to have an effect on [one's] environment as well as to attain valued outcomes within it"), and relatedness (the need "to feel connected to others") (Deci and Ryan, 2000: 231). As Katherine Bartlett (2009) summarizes in the *Virginia Law Review*:

> [T]hreat and confrontation about race and gender bias, which people do not want to possess or exhibit, may inadvertently provoke shame, guilt, and resentment, which lead to avoidance and resistance, and ultimately to more stereotyping. In other words, pressure and threat will often deepen bias rather than correct it. Positive strategies that affirm people's good intentions, in contrast, engage people constructively in defining their better, nondiscriminatory selves and aligning their conduct accordingly. (1901)

Knowledge of implicit biases over which they have no awareness or control—and the prospect of being negatively judged on the basis of it—can decrease people's experiences of autonomy and competence.[31] And feeling that they may be blamed or viewed as racist or sexist, unable to trust that they can proceed on others' good faith and the benefit of the doubt, undoubtedly prevents feelings of relatedness. All of this hampers the *internalization* of egalitarian norms, which, as Devine and Plant (1998) showed, would help reduce implicit bias. Encouraging the internalization of egalitarian norms requires putting people in environments where they can feel autonomous, competent, and related to others. This is far more easily achieved in a climate where appraisal-based responses are avoided—that is, where blaming, shaming, and resentment are avoided. Non-appraising responses, by contrast, serve to enhance feelings of autonomy and competence by pointing to ways in which agents can actively make amends for their mistakes and improve their efforts to attain egalitarian ideals.

An environment where appraisal-based responses are common can also lead to more subtle and unexpected harms. For example, it appears that widespread fear or aversion to being perceived (or perceiving oneself) as racist can cause teachers to be less effective. Harber, Stafford, and Kennedy (2010), for instance, demonstrate the existence of a "positive feedback bias," in which teachers whose commitments to egalitarianism are challenged subsequently grade the work of minority students less strictly than that of White students. (The effect does not occur when their egalitarianism is affirmed.) Such skewed grading, however, has negative consequences. When perceived by minority students, overly positive feedback can serve to erode their trust in the mainstream educational system, alienate or underchallenge them, and deprive them of useful feedback. But Croft and Schmader (2012) found that the feedback bias was activated primarily among externally motivated egalitarian instructors who were "concerned about appearing racist," and Harber et al.'s (2012) most recent study found that, interestingly enough, positive feedback bias toward Black students decreased when teachers were in a more socially supportive environment, where measures of social support included the degree to which they felt their colleagues were friendly and supportive. Moreover, Norton et al.'s (2013) research on "racial paralysis" found that Whites' concerns about appearing racially biased led them to make significant efforts to avoid making cross-race judgments, while

[31] This does not mean that we should conceal the existence of implicit biases, but that education and training should always include, for example, constructive strategies for mitigating their effects. Such education would also do well that emphasize that *everyone* harbors some such set of biases and hence faces the same problem, so to be informed of one's bias is not to be singled out as a "bad person."

Plant and Devine's (2003) and Plant and Butz's (2006) studies on interracial anxiety showed that the fear of appearing biased led White people to avoid interracial interactions—all of which can only serve to strengthen implicit bias. Notably, research in self-affirmation theory has demonstrated that people who feel valued and affirmed (in unrelated domains) subsequently show greater openness to potentially threatening information and more willingness to compromise, leading to improved intergroup relationships (see Cohen and Sherman 2014 for a review). Feeling socially supported and promoting positive intergroup relationships thus require trusting climates in which people's motives and character are not under suspicion, and this can almost certainly be better achieved by eschewing appraisal-based responses in favor of non-appraising ones.

This is where I part company with Jules Holroyd (2012), who argues against such pragmatic reasons by claiming that taking ourselves to be liable to blame for being influenced by implicit biases may have beneficial effects even if no one is in an epistemic position to actually engage in blaming. Holroyd writes:

> Classifying certain actions as prohibited, for which individuals are liable to blame, can have numerous important effects, including: strengthening norms against so acting; encouraging individuals to self-monitor; and leading us to change our expectations of the steps others might take in monitoring their own behavior... [N]ote that [these effects] do not depend upon us being able to in fact engage in blaming, although some of them might encourage us to challenge others' decisions and provide careful justification for them. (2012: 300)

The whole problem with implicit biases, however, is that *being influenced by an implicit bias* is not the sort of "action" or "behavior" that it is useful to prohibit. Discriminatory behavior itself—the thing that can be called an action or a way of acting—has long been prohibited. Strengthened norms against being unconsciously influenced by implicit bias, if it even makes sense for agents to follow norms that by definition they cannot consciously try to follow, is more likely to have the effect of threatening people's conceptions of themselves as committed to egalitarianism. Self-monitoring and monitoring of others, on the other hand, can be required and expected as part of holding people accountable, without inducing the further requirement that they feel blameworthy. Again, feeling liable to these appraisal-based responses is probably counterproductive.[32]

[32] There is a fascinating psychological literature on the difference between shame and guilt. The dominant theory of the distinction is that "shame involves a negative evaluation of the global self; guilt involves a negative evaluation of a specific behavior." Moreover, guilt is found to be far more adaptive and productive than shame: shame motivates efforts to hide or escape the problem, while guilt is other-focused and motivates reparative actions (Tangney, Stuewig, and Mashek, 2007: 349). This suggests that guilt is not simply self-blame, as philosophers have commonly taken it to be, and that insofar as it is non-appraising it may still be warranted.

Before concluding, I should emphasize that mine is not a purely consequentialist account of moral responsibility, on which appraisal-based responses are justified or unjustified purely in virtue of good or bad consequences. Recall my two-pronged argument that internally motivated egalitarians often do not, for moral reasons, deserve appraisal-based responses and externally motivated egalitarians should not, for pragmatic reasons, be confronted with them. Desert still determines the aptness of appraisal-based responses, but the formidable epistemic barriers to establishing it means that appraisal-based responses will often be morally unjustified.[33] (Thus desert is a necessary but not sufficient criterion on my view. A person must be attributively responsible, and we must be able to confidently know that, in order for appraisal-based responses, e.g. judgments of blameworthiness, to be justified.[34]) In other cases, appraisal-based responses may be truly warranted, but the moral reasons for engaging in them, e.g. by overtly blaming, are outweighed by pragmatic considerations. I should emphasize that accountability too goes beyond merely procuring good consequences: the shape of a given society's social arrangements is also a representation of its members' relations to one another, and is thus capable of expressing (or failing to express) mutual respect and equal concern. Systems of accountability do not themselves rest on purely consequentialist grounds of justification.

Let me caution also that my recommended solution should be understood as a general policy, one that can justifiably be overridden. The appropriateness of various practices of critical moral response is always sensitive to context, especially the contexts of particular relationships, and the nature of these relationships may license deviations from my main thesis.[35] For example, in the case of

[33] To mention just one more such barrier, it may be difficult to discern the difference between, say, "unconscious racism" and implicit racial bias. I take it that if such a distinction can be made tenable, the former would ground attributability even while the latter might not. The difference would lie in the robustness and sophistication of the subpersonal machinery: unlike unconscious racism, racial implicit bias as measured by the Implicit Association Test, for instance, might be a much thinner cognitive association (as, say, in the case of an association between "salt" and "pepper") that is not hooked up to other elements of a person's psychological economy. I am grateful to Nathaniel Coleman, Janine Jones, Megan Mitchell, and Chandra Sripada for discussions on this point.

[34] Cf. Young (2011): "When applying this model of responsibility [as guilt, blame, or liability], there should be clear rules of evidence, not only for demonstrating the causal connection between this agent and a harm, but also for evaluating the intentions, motives, and consequences of the actions."

[35] For this reason it might be more apt to conceive of the aptness of critical moral responses as relative, rather than absolute: for example, a person P is blameworthy by some set of others S in situation C (or in the context of relationship R or with social identity I or in background conditions B), not blameworthy *tout court*. I do not have the space to defend such a view here, but see Springer (2013), Kutz (2000), and Calhoun (1989).

self-relation, it might be appropriate for me to feel self-blame. There my epistemic situation with regard to myself is different from that with regard to others, and I may well be in a position to know that I have not done all that could be reasonably expected of me to combat bias.[36] Similarly, a close friend with whom I have had a long history might be in a position to chastise me for slipping up yet again. In this case, however, it may be that I have violated the second condition by not taking enough due care with an implicit bias that I could reasonably be expected to take, and my friend has good enough epistemic grounds on which to thus attribute the biased behavior to me. Or, it may be that her chastisement—while it has the appearance of an appraisal-based response—does not express the negative appraisals that are characteristic of blame; here it would function more as a mechanism of accountability, a way of indicating where I have fallen short and reminding me to be more cautious in the future. In the case of public figures, too, the situation might be different insofar as there is a distinction between a person and her public persona, and in light of ramifications for public discourse, role modeling, and so on; while refraining from appraisal-based responses will likely be all the more important here, it may be that particularly egregious cases deserve to be called out. All of these possibilities are consistent with my view.

Finally, where there are only pragmatic reasons to refrain from appraisal-based responses, as in the case of externally-motivated egalitarians, these may be trumped by other, stronger reasons for engaging in them. For example, if there are cases in which it is critical for a person's being able to heal from the trauma of being subjected to biased behavior that she name it for what it is, then even overt blame may be justified. While confrontation is generally counterproductive and should be avoided—which is easier where there are well-established, effective structures of accountability!—we cannot demand of the victims of bias

[36] Smith (2004: 347) has made the pragmatic argument that viewing unendorsed implicit attitudes as non-attributable to oneself may turn out to be a self-fulfilling prophecy, in that it can blind, demotivate, or otherwise prevent a person from trying to change those parts of herself. To the extent that this is true, it would be a reason to allow appraisal-based responses toward oneself. However, Young (2011) argues just the opposite, that feeling blameworthy or guilty represents a "self-indulgence" that is "unproductive" because it focuses attention on oneself rather than what needs to change in the world. Thus the appropriateness of blaming oneself will probably need to be assessed on a case-by-case basis. While empirical research (e.g. Monteith et al., 2002) has demonstrated that negative self-directed affect can play an important role in the self-regulation of prejudice, it has hitherto failed to distinguish between specific types, e.g. shame versus guilt, which may be significantly different (as per fn. 32). It is also worth noting that we might be able to avoid the paralyzing effect Smith is concerned about by holding ourselves *accountable* for implicit biases, where accountability consists precisely in undertaking efforts to change. We could undertake these efforts without feeling the shame that comes of viewing our implicit biases as genuinely attributable to ourselves.

that they should, in addition to suffering the harms of biased behavior, bear the further burden of absorbing such damage to their self-respect.[37] It is also possible, though the calculation of risks is an empirical matter difficult to ascertain in particular cases, that there may be long-term effects that could outweigh immediate negative consequences. One hopes, for instance, that a person who gets called out on their (unconscious) racism might eventually come to understand themselves partly in virtue of having undergone that experience. As I mentioned, however, this is risky, and it seems on balance better pragmatically to avoid it as a strategy. But the existence of this possibility means that I certainly do not deny the need for there to be a wide range of approaches "out there" in the world, including both appraisal-based and non-appraising responses, in order to address all levels of receptiveness. A person who is initially best served by a non-appraising, largely educational response may develop to a point where they may be receptive to certain kinds of appraisal-based response, but this will usually require supportive, low-threat conditions.

6 Conclusion

For pragmatic reasons alone, then, I might have been able to argue for the view that we ought to hold people accountable but not attributively responsible for implicit bias. But I do not want to lose sight of the moral reasons for avoiding appraisal-based responses. In my view, pragmatic reasons are more often than not tied up with moral reasons; a person's psychological recalcitrance regarding critical moral response may itself be a reaction against others' failure to acknowledge important aspects of what it is for her to be and to conceive of herself as an imperfect moral agent struggling in an unjust world, or it may indicate a troubling lack of moral community and relations of mutual respect between criticizer and criticized. Autonomy, competence, and relatedness are important for internalizing and realizing social norms because these are precisely the conditions that express respect for people's experiences of themselves and others as efficacious moral agents responding to reasons as best they can. As we know, implicit biases act on us in ways that undermine the exercise of our rational capacities for self-reflective deliberation, choice, and endorsement—the things that make it possible for us to be morally responsible beings at all. They can be utterly invisible to us. Thus, in respecting people as morally responsible agents, we should pay attention to what it is like for an agent trying to act rightly within

[37] I am indebted to Kristie Dotson for this point.

the limits of what is visible from her practical point of view.[38] But, as Angela Smith (2005) has argued, being held responsible is not only a burden but a privilege: an expression of respect for a person's status as moral agent. Refusing to hold people responsible relegates them to the category of beings with whom we cannot have fully developed moral relations. This is why it remains vitally important that we still hold people accountable for implicit bias: that we view them as agents whose actions express their social relationships to us. After all, what makes it necessary for us to hold each other responsible in the first place are the *social* needs based on relationships within the moral community and what is required to fashion and uphold acceptable forms of those relationships. The trick is to figure out how to be sensitive to the complexities of human action and the many ways human agents fall short—all the while still holding them accountable when they do. I hope that this chapter has contributed to that task.

Acknowledgments

I am grateful for comments on earlier drafts of this paper from Elizabeth Anderson, Sarah Buss, Kristie Dotson, Neil Levy, Ishani Maitra, Angela Smith, and the editors of this volume, as well as feedback from participants at the 2012 Implicit Bias and Philosophy Project Workshop and the 2013 Eastern APA Division Meeting.

References

Alicke, M. D., Buckingham, J., Zell, E., and Davis, T. (2008). "Culpable control and counterfactual reasoning in the psychology of blame." *Personality and Social Psychology Bulletin* 34: 1371–81.

Arpaly, N. (2004). *Unprincipled Virtue: An Inquiry Into Moral Agency*. New York, NY: Oxford University Press.

Baier, K. (1987). "Moral and Legal Responsibility." In Siegler, M., Toulmin, S., Zimring, F., and Schaffner, K. (eds), *Medical Innovation and Bad Outcomes: Legal, Social, and Ethical Outcomes*. Ann Arbor, MI: Health Administration Press: 101–30.

Baron, A. S. and Banaji, M. R. (2006). "The development of implicit biases: Evidence of race evaluations from ages 6 and 10 and adulthood." *Psychological Science* 17(1): 53–8.

Baron, M. (1987). "Kantian ethics and supererogation." *The Journal of Philosophy* 84(5): 237–62.

Bartlett, K. T. (2009). "Making good on good intentions: The critical role of motivation in reducing implicit workplace discrimination." *Virginia Law Review* 95: 1893–972.

Bertrand, M. and Mullainathan, S. (2004). "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination." *The American Economic Review* 94(4): 991–1013.

[38] Cf. Korsgaard (1992: 323–4).

Blum, L. (2004). “Stereotypes and stereotyping: A moral analysis.” *Philosophical Papers* 33(3): 251–89.

Calhoun, C. (1989). “Responsibility and reproach.” *Ethics* 99(2): 389–406.

Cohen, G. L. and Sherman, D. K. (2014). “The psychology of change: Self-affirmation and social psychological intervention.” *Annual Review of Psychology* 65: 333–71.

Croft, A. and Schmader, T. (2012). “The feedback withholding bias: Minority students do not receive critical feedback from evaluators concerned about appearing racist.” *Journal of Experimental Social Psychology* 48(5): 1139–44.

Czopp, A. M., Monteith, M. J., and Mark, A. Y. (2006). “Standing up for a change: Reducing bias through interpersonal confrontation.” *Journal of Personality and Social Psychology* 90(5): 784–803.

Dasgupta, N. and Greenwald, A. G. (2001). “On the malleability of automatic biases: Combating automatic prejudice with images of admired and disliked individuals.” *Journal of Personality and Social Psychology* 81(5): 800–14.

Deci, E. L, and Ryan, R. M. (2000). “The ‘what’ and ‘why’ of goal pursuits: Human needs and the self-determination of behavior.” *Psychological Inquiry* 11: 227–68.

Devine, P. G. and Plant, E. A. (1998). “Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology* 75(3): 811–32.

Fischer, J. M. and Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge, MA: Cambridge University Press.

Fischer, J. M. and Tognazzini, N. A. (2011). “The physiognomy of responsibility.” *Philosophy and Phenomenological Research* 82(2): 381–417.

Frankfurt, H. (1988). *The Importance of What We Care about: Philosophical Essays*. Cambridge, MA: Cambridge University Press.

Fricker, M. (2007). *Epistemic Injustice*. Oxford: Oxford University Press.

Gawronski, B., Hofmann, W., and Wilbur, C. J. (2006). “Are ‘implicit’ attitudes unconscious?” *Consciousness and Cognition* 15: 485–99.

Goodin, R. (1995). *Utilitarianism as a Public Philosophy*. Cambridge: Cambridge University Press.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., and Banaji, M. (2009). “Understanding and using the Implicit Association Test III: Meta-analysis of predictive validity.” *Journal of Personality and Social Psychology*, 97(1): 17–41.

Harber, K. D. et al. (2012). “Students’ race and teachers’ social support affect the positive feedback bias in public schools.” *Journal of Educational Psychology* 104(4): 1149–61.

Harber, K. D., Stafford, R., and Kennedy, K. A. (2010). “The positive feedback bias as a response to self-image threat.” *British Journal of Social Psychology* 49: 207–18.

Hieronymi, P. (2004). “The force and fairness of blame.” *Philosophical Perspectives* 18(1): 115–48.

Holroyd, J. (2012). “Responsibility for implicit bias.” *Journal of Social Philosophy* 43(3): 274–306.

Kelly, E. (2002). “Doing without desert.” *Pacific Philosophical Quarterly* 83(2): 180–205.

Korsgaard, C. (1992). “Creating the kingdom of ends: Reciprocity and responsibility in personal relations.” *Philosophical Perspectives* 6: 305–32.

Korsgaard, C. (2009). *Self-Constitution: Agency, Identity, and Integrity*. New York, NY: Oxford University Press.

Kutz, C. (2000). Complicity: Law and Ethics for a Collective Age. Cambridge: Cambridge University Press.

Langton, R. (2001). "Virtues of Resentment." *Utilitas* 13(2): 255–62.

Legault, L., Gutsell, J. N., and Inzlicht, M. (2011). "Ironic effects of antiprejudice messages: How motivational interventions can reduce (but also increase) prejudice." *Psychological Science* 22(12): 1472–7.

Levy, N. (2014). "Consciousness, implicit attitudes and moral responsibility." *Noûs* 48(1): 21–40.

Malle, B. F., Guglielmo, S., and Monroe, A. E. (2014). "A theory of blame." *Psychological Inquiry* 25(2): 147–86.

Monteith, M. J., Ashburn-Nardo, L., Voils, C. I., and Czopp, A. M. (2002). "Putting the brakes on prejudice: On the development and operation of cues for control." *Journal of Personality and Social Psychology* 83(5): 1029–50.

Moody-Adams, M. M. (1994). "Culture, Responsibility, and Affected Ignorance." *Ethics* 104(2): 291–309.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M., and Handelsman, J. (2012). "Science faculty's subtle gender biases favor male students." *Proceedings of the National Academy of Sciences* 109: 16474–9.

Norton, M. I., Mason, M. F., Vandello, J. A., Biga, A., and Dyer, R. (2013). "An fMRI investigation of racial paralysis." *Social Cognitive and Affective Neuroscience* 8: 387–93.

Olson, M. A. and Fazio, R. H. (2006). "Reducing automatically activated racial prejudice through implicit evaluative conditioning." *Personality and Social Psychology Bulletin* 32(4): 421–33.

Oshana, M. (1997). "Ascriptions of responsibility." *American Philosophical Quarterly*, 34: 71–83.

Pearson, A. R., Dovidio, J. F., and Gaertner, S. L. (2009). "The nature of contemporary prejudice: Insights from aversive racism." *Social and Personality Psychology Compass* 3.

Plant, E. A. and Butz, D. A. (2006). "The causes and consequences of an avoidance-focus for interracial interactions." *Personality and Social Psychology Bulletin* 32: 833–46.

Plant, E. A. and Devine, P. G. (2001). "Responses to other-imposed pro-black pressure: Acceptance or backlash? *Journal of Experimental Social Psychology* 37: 486–501.

Plant, E. A. and Devine, P. G. (2003). "The antecedents and implications of interracial anxiety." *Personality and Social Psychology Bulletin* 29(6): 790–801.

Rawls, J. (1999). *A Theory of Justice*. Cambridge, MA: Belknap Press.

Rowe, M. (2008). "Micro-affirmations and micro-inequities. *Journal of the International Ombudsman Association* 1(1): 45–8.

Saul, J. (2013). "Implicit bias, stereotype threat, and women in philosophy." In Jenkins F. and Hutchison, K. (eds.), *Women in Philosophy: What Needs to Change?* Oxford: Oxford University Press.

Scanlon, T. M. (1998). *What We Owe Each Other*. Cambridge, MA: Belknap Press.

Shoemaker, D. (2011). "Attributability, answerability, and accountability: Toward a wider theory of moral responsibility." *Ethics* 121(3): 602–32.

Smiley, M. (1992). *Moral Responsibility and the Boundaries of Community*. Chicago, IL: University of Chicago Press.

Smith, A. (2004). "Conflicting attitudes, moral agency, and conceptions of the self," *Philosophical Topics* 32(1–2): 331–52.

Smith, A. (2005). "Responsibility for attitudes: Activity and passivity in mental life." *Ethics* 115(2): 236–71.

Smith, A. (2007). "On being responsible and holding responsible." *Journal of Ethics* 11(4): 465–84.

Smith, A. (2012). "Attributability, answerability, and accountability: In defense of a unified account." *Ethics* 122(3): 575–89.

Springer, E. (2013). *Communicating Moral Concern*. Cambridge, MA: MIT Press.

Sripada, C. (2015). "Self-expression: a deep self theory of moral responsibility." *Philosophical Studies*: 1–30.

Strawson, P. F. (1962). "Freedom and resentment." *Proceedings of the British Academy* 48: 1–25.

Sue, D. W. (2010). *Microaggressions in Everyday Life: Race, Gender, and Sexual Orientation*. New York, NY: Wiley.

Tangney, J. P., Stuewig, J., and Mashek, D. J. (2007). "Moral emotions and moral behavior." *Annual Review of Psychology* 58: 345–72.

Velleman, D. (1992). "What happens when someone acts?" *Mind* 101: 461–81.

Wallace, R. J. (2004). *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.

Watson, G. (2004). *Agency and Answerability: Selected Essays*. New York, NY: Oxford University Press.

Wolf, S. (1990). *Freedom Within Reason*. Oxford: Oxford University Press.

Young, I. M. (2011). *Responsibility for Justice*. New York, NY: Oxford University Press.

1.4

Stereotypes and Prejudices: Whose Responsibility?

Indirect Personal Responsibility for Implicit Biases

Maureen Sie and Nicole van Voorst Vader-Bours

1 Introduction

This chapter is a philosophical paper on how to understand the role of stereotypes and prejudices in human life, their harms and benefits, and our, what we will call, "indirect personal responsibility" with respect to the harms. We discuss the harms and benefits in detail because we believe it is important, for reasons that will become clear later, to recognize that we cannot take for granted that we all agree on (1) what these harms and benefits are, (2) that the harms can and should be prevented, and (3) that it is clear who bears the responsibility for perpetrating harm. It is important to keep in mind that in this chapter we are not primarily interested in clear instances of discrimination: passing someone over for promotion on the sole ground that she is a woman, for example. Rather, we are interested in those many instances in which stereotypes and prejudices influence our interactions in more elusive ways, often even without awareness of the harm done as would be the case when we act under the influence of implicit biases. This is also why we focus on "indirect personal responsibility": we are mainly interested in the responsibility we might bear for stereotypical and/or prejudiced behavior that causes harm, even in cases where conscious control and awareness are absent.

The starting point for our investigation is an insight gained from Kelly and Roedder's overview of empirical work on racial cognition in which they connect stereotypes and prejudices with implicit biases, and discuss what consequences the role of implicit race biases may have for moral theory (Kelly and Roedder, 2008). The main characteristic of implicit biases is that they often escape our conscious control and awareness as we act: they bias what we do whilst remaining implicit.[1] One of the questions Kelly and Roedder pose is whether we have an epistemic "duty" to correct for implicit biases in our daily undertakings. They suggest that when making more considered, deliberative judgments we may have the epistemic duty to compensate for the impact of implicit biases whenever we suspect that we are affected by them (535). We argue that this responsibility is best understood as an indirect one, deriving from our place as members of a collective that sustains the implicit biases underpinning many of the stereotypes and prejudices in our culture.

Our argument, roughly outlined, is the following:

1. Stereotypes and prejudices (hereafter S&Ps) stem from features of our general cognitive make-up that enables us to function in a socially complex world such as ours.
2. S&Ps are social constructs that are formed, shared, and upheld by a collective of individuals.
3. Some S&Ps are harmful.
4. When this harm is recognized by a collective of individuals (a society, group, organization, and so on) and (a) that collective fails to discontinue the effects of the harmful S&Ps, and when (b) implicit biases (from now on IBs) are the probable cause of this failure, then members of that collective (c) have a responsibility to compensate or correct for the impact of IBs on their behavior if they are in the position to do so.
5. This responsibility to correct for the impact of IBs is derived in a straightforward way from the commitment to discontinue the efficacy of the harmful S&Ps.

As said, we call this an argument for our indirect personal responsibility for S&Ps. The reason for this is that it is an argument that is independent of our answer to the question whether we can control the influence of IBs on our behavior or whether we can directly counteract or avoid the influence of IBs altogether.[2] When we control/are able to counteract the influence of IBs, we are also directly responsible for acting under the influence of IBs. For the purposes of

[1] We can become aware that that we foster a specific implicit bias. For example, if we have taken an IAT (see the Introduction in Volume 1 for an explanation of the IAT). When we are aware of our implicit biases, we might also be able to control them, as we discuss in Section 4.

[2] Whether we can help acting in that way is the topic of lively debate at the moment. See, for example, Kelly and Roedder (2008: 532); Saul (2013); Holroyd (2012).

this chapter, we leave that issue unaddressed. Another point that we do not discuss is whether we should be committed to discontinuing harmful S&Ps, though we give some reasons for such a commitment in our discussion concerning the harm they do. A last caveat: we do not elaborate on the sense in or amount to which we should compensate or correct for the impact of IBs, or what exactly this implies. We do, however, elaborate on some potential debiasing measures in Section 4. Our main point is to establish indirect responsibility and explain why it is attractive to think in such terms with respect to our responsibility vis-à-vis harmful S&Ps. As we will explain, we believe that the matter of what harm is done by so-called harmful S&Ps is a complicated issue, even though it often quite clear who is harmed. Related to this, we also believe that answering the question of who is to blame for the harm done is not straightforward, even though there is much we can do to contribute to the collective enterprise of discontinuing harmful S&Ps.

We believe our argument for indirect responsibility has some advantages. It explains why harmful S&P behavior is persistent and hard to change, even when the majority of people are willing to change it. Our argument also makes clear that we need to evaluate our attitudes and behavior vis-à-vis S&Ps in a more fine-grained manner and to acknowledge the harm that is inflicted without oversimplifying who, if anyone, is to blame for it. With respect to this last point, we think that a focus on harm rather than on perpetrators might stimulate a more open approach regarding understanding and discussion of harmful S&Ps (which affect some groups more than others), and what we can do about them. We start in Section 2 by explaining what S&Ps are and why we think they are deeply embedded in our everyday interactions with one another. We continue by explaining why certain S&Ps are harmful and why we have reason to want them to disappear or change. We emphasize the fact that S&Ps are collectively constructed and upheld. As a consequence, we argue, with respect to S&P behavior, that the notion of individual "perpetrators" and "victims" is complex.

In Section 3 we address the role and function of IBs in S&P behavior and develop an argument in defense of indirect responsibility for persistently disavowed S&P behavior. In this section we also explain the advantages of talking in terms of indirect responsibility.

In Section 4 we present some of the literature concerning debiasing measures we could take that can contribute to the discontinuation of harmful S&Ps inasfar as they are caused by IBs. In addition, we discuss an example in which we try to identify factors that might have been at play in activating stereotypes and show what we could learn from such an enterprise. This discussion illustrates that, even if we believe that there is nothing we can do to counteract the influence of certain

S&Ps on our behavior since IBs escape our conscious control and awareness, scrutinizing our behavior in a certain manner might still be extremely fruitful.

2 Stereotypes, Prejudices, and their Harm

In everyday life we confront a continuum of possible choices with regard to evaluating others as a basis for our assessments and subsequent acts (and omissions). At one end of the continuum, we rely on general characteristics and categories; at the other end, we individuate people, focusing on their personal characteristics. Relying on S&Ps is quick and effortless, aiding in general categorizations, but may result in unfair, perhaps even unjustifiable, judgments and subsequent harmful actions. In this section we elaborate on what stereotypes and prejudices are and how we should understand and evaluate the harm they might do. We emphasize those aspects that are important to our argument for indirect responsibility as outlined in the introduction.

2.1 Stereotypes and prejudices

The term “stereotype” was coined by Lippmann to describe the (simplified and culturally informed) “little pictures” we carry around inside our heads about various “types” of people (Lippmann, 1922: 63). Stereotypes are commonly understood to be generalizations or, in terms of Valian (2005), “schemas.” These generalizations are simplified and standardized images that ascribe specific (psychological or other) characteristics to all members of a specific social category, without individuating between them.[3] They are (some of) the cognitive underpinnings of prejudice. Prejudices are the often affective, evaluative applications of triggered stereotypes. Hereafter we treat stereotypes and prejudices as knit so closely together that it makes sense to just speak of S&P behavior.

In the early days of thinking about S&Ps, Allport (1954) observed that orderly living depends on categorizations (stereotypes) that, once formed, are the basis for our everyday prejudgment that, he argued, can easily become prejudices, but “only if they are not reversible when exposed to new knowledge” (9). Also, people form in-groups, and reject out-groups regularly. We agree with both observations and think that there are several other aspects which contemporary findings have brought to the surface that help explain the nature of S&Ps. In our explanation of what S&P behavior is, we focus on the following aspects: (a) the collective nature of stereotype formation, upkeep, and discontinuation, (b) the judgment-guiding, action-guiding and evaluative character of S&Ps, (c) the multiplicity of traits that

[3] See, for example, Allport (1954); Jost and Banaji (1994); and many others.

can be simultaneously attributed (both negatively and positively valenced ones), and (d) the often almost automatic, uncritical and non-reflective assumption that a stereotypical trait "fits" the stereotyped target—without verifying whether the trait attribution and/or the lack of individuation is warranted.

To begin with the collective nature: crucial to stereotypes is that they are socially shared cultural constructs. They are "shared group beliefs" that are "formed in line with accepted views or norms of social groups."[4] A mere generalization is bound to remain an idiosyncrasy, an individual construct, as long as others do not share it. Let us clarify. Suppose, for example, that I sincerely thought that people with an apple-shaped body are kind, whereas people with a pear-shaped one are jealous.[5] Although I made a generalized observation, we would not call it a stereotype because others do not share my idea. This would change once my individual generalization is shared among a substantial group of people; for example, all my peers or all people with the same gender as mine. The thought that body shape is related to a particular character trait is now an association "we" share, for example, within the considerable group of my peers or the larger group of people with my gender. In order for it to become a stereotype in the usual sense, it needs to acquire a substantial "life of its own." This has not been the case with the association of jealousy and kindness with body shape, but in some medical and lay circles it is thought that "apple-shaped" people run more health risks than "pear-shaped" ones because of the way their body fat is distributed. The validity of this automatic association has only recently been refuted, but will likely remain a stereotype among those circles for as long as the medical evidence has not convinced them.[6]

Due to their nature as collective construals, stereotypes will only "survive" if they are collectively upheld; otherwise they will wither away. One of the stereotypes that dissolved is that of the Irish as "dirty, drunken, incompetent, brawling slum-dwellers" (mid-1800s USA).[7] Some stereotypes even change into their opposite. In the nineteenth century, illegal immigration in the USA was a problem with a Chinese face: the Chinese were considered despicable, intellectually inferior

[4] See McGarty, Yzerbyt, and Spears (2002: 3). Usually, three principles are identified that all characterizations of stereotypes share to a greater or lesser degree: (1) culturally shared; (2) aids to explanation; (3) energy-saving devices.

[5] "Apple-shaped" refers to people whose weight is concentrated around the abdomen, whereas "pear-shaped" refers to those who carry weight more in the buttocks, hips, and thighs.

[6] Due to several rounds of reviews, we learned that many have not heard of this particular stereotype, so for an article combating it, see <http://www.sciencedaily.com/releases/2013/01/130110161350.html>.

[7] Kunda (1999: 391).

workers. Nowadays, the manner of how they are characterized has changed dramatically to the opposite.

Let us summarize. For a stereotype to "survive" it needs to be regularly employed and affirmed by a sufficiently large group. Only then will we attribute this specific trait to members of the targeted group so that our everyday associations will remain imbued with this stereotype. Only then will the younger generation (automatically) come to share these stereotypes because they are part of the pool of social cognition in the society of which they are a part. Otherwise, the stereotype will wither away. For an individual to cease to hold a stereotype, the associative, cognitive, and affective links in their mind between traits and social groups need to be altered or severed. But for a stereotype itself to disappear, there must be a collection of individuals who no longer employs it. We will come back to this later.

The second aspect that is important to the argument of this chapter is the judgment-guiding, action-guiding, and evaluative nature of S&Ps. Stereotypes are commonly understood as relatively "cold" cognitions that are predictive with regard to judgment, impression formation, and action. In contrast, prejudices are generally characterized as "hot," often affective, evaluations that are predictive for interpersonal preferences and social distance (e.g. seating distance).[8] We acknowledge this cognitive–evaluative distinction. However, it seems that (1) both processes are typically congruent and work together to facilitate coordinated responses,[9] and that (2) some stereotypes are likely to have a less cognitive and more affective and evaluative nature than others. Racial stereotypes might be "hotter" and more affectively laden than other types of social and non-social attitudes (Dovidio and Gaertner, 1996). We do not see a strict dichotomy between cognition and affect, and think it is often difficult to pry apart cognitive elements from the evaluative ones, even in less "hot" cases. Consider for example the phrase "silly teenaged girls." The very idea of linking a specific trait (e.g. "silly") with a category of people ("teenaged girls") bears an evaluative aspect in it. In a similar vein, prejudices have a cognitive element to them, if only because they are informed by stereotypes. To put it differently, stereotypes are the predominantly cognitive underpinnings of prejudice; prejudices are the mainly evaluative application of triggered stereotypes, and often the twain shall intermingle. What is important for the purposes of this chapter is the fact that our

[8] See, for example, Greenwald and Banaji (1995); Amodio and Devine (2006). The predicate "hot" does not necessarily indicate anger, aggression, violence, or the like, but rather an emotional evaluation of any kind (e.g. dislike, antipathy, fear, and so on).

[9] See, for example, Amodio and Devine (2006).

interactions (actions, judgments, and evaluations) are thoroughly influenced by a mixture of S&Ps. For example, when due to S&Ps we treat teenaged girls in a slightly condescending way and avoid seats next to African American males on public transport, it is hard to separate the cognitive aspects from the affective ones.

The third aspect we identified as important to our overall argument is that stereotypes reflect the attribution of one or more traits that are believed to be characteristic of the members of a social category that often differ in valence. They encompass the totality of information that people have in mind when they think about various groups "as a group" rather than as individuals.[10] These traits can be positive ("fashion models are physically attractive"), negative ("they are empty-headed") or neutral ("they are young"). People often simultaneously attribute a variety of negative, positive, and neutral attributes to a group. By neutral valence we mean stereotypes that are, or seem to be, predominantly descriptive in the attributed trait without reflecting any evaluative judgment. For example, stereotypes that lawyers are assertive or Dutch people are tall have neutral valence.[11] Another example is that when we see a toddler dressed in blue or pink, we almost automatically assume it is respectively a boy or a girl (we return to this example later).

S&Ps are a complex mix of cognitive and affective components with different valences that guide our judgments, actions, and evaluations. It is far from easy to disentangle the cognitive from the affective components, let alone change them. Moreover, the triggering of S&P behavior is context-dependent. If the model from the previous example, for instance, attends a dinner where most guests are academics, the stereotype "empty-headed" may be triggered; if the dinner celebrates her parents' wedding anniversary, it is more likely that "young" comes to mind. These contextual aspects are also difficult to identify because of the automaticity of most S&P behavior. This brings us to the final aspect that we mentioned.

We are inclined to uncritically, non-reflectively assume that certain traits "fit" the stereotyped target without verifying whether the trait attribution and/or the lack of individuation is warranted. Take again the "male and female colors" example. In earlier times, male and female infants alike used to be clothed in

[10] See, for example, Macrae and Bodenhausen (2001).

[11] Tallness seems neutral. However, research findings suggests that there is a significant association between height and perceived leadership for which an evolutionary explanation can be given (Murray and Schmitz, 2011). Thus, although the description in itself might be neutrally valenced, it actually evokes associations with a positive valence (i.e. ability).

white dresses (that is, dresses in the sense of a “gown” rather than a typically female piece of clothing). From the beginning of the twentieth century, pink became associated with boys (red symbolizing “zeal and courage”) and light blue with girls (blue symbolizing “faith and constancy”). Only since the mid-twentieth century has the opposite association—pink for girls and soft blue for boys—become the fixed gender label.[12] Now imagine that we come across a painting of a child wearing soft blue. Because of the strong association in the modern western world between pink and girls and soft blue and boys, we will probably infer that the child is a boy. When the painting is from the early twentieth century, we are likely to be mistaken. It is only when it is pointed out to us that at the time the portrait was painted blue signaled “girl” that we become aware that our idea that the child is a boy was based upon the color of the outfit of the child. We did not make that inference explicitly; on the contrary, we “just” automatically associated blue with boys. It is this automaticity that makes S&Ps efficient but also difficult to change and/or counteract.

In many cases this automaticity is unproblematic; moreover, without it S&Ps would not facilitate speedy processing of information, often to our benefit. This is also true for stereotypes with a clear one-sided positive or negative valence; for example, the stereotype that police officers are authoritarian figures who are not to be trusted, or the opposite stereotype that they are trustworthy individuals to whom one can turn when in need. Relying on those stereotypes automatically and without prior thought can save our lives.

Of course, S&Ps, by their very nature, can also “misfire”—figuratively speaking. There are many Dutch people who are not tall at all, and even in societies in which police officers tend to be your best friend there are some individual officers who are not to be trusted. When we expect a tall person to turn up at our meeting with a Dutch philosopher, we might not spot her because she is not tall at all. Also, some S&Ps misfire in the sense that they are misattributions that do not apply in general or in a specific instance. As indicated earlier, recently it has been found that pear-shaped people are not, generally speaking, healthier than apple-shaped people. But what exactly makes certain S&Ps harmful and others not, or, perhaps more adequately put, what makes certain S&P behavior more harmful than other S&P behavior? Is it related to how adequate they are and how often they “misfire”? Is our resistance to certain S&Ps that they are not adequate or false and is this why they are harmful? Let us turn to discussing the harm of S&P behavior.

[12] According to sociologist Jo Paoletti; see Fine (2006: 207, 208, and 283 endnote 1).

2.2 Harmful S&P behavior

When reflecting on S&P behavior, we might think that it is their negative valence that causes the harm. When we acknowledge the function S&Ps serve by facilitating speedy processing of information, as outlined in the previous section, this negative valence might sometimes be the price attached to the advantages they bring us, as in the example of the police-officer. Of course, even neutrally valenced stereotypes might cause harm. An infertile woman, who is constantly seen as 'able to have children', for example, might be hurt by questions that remind her of her infertility. Hence, the mere fact that S&Ps ignore individual differences even when relevant can cause harm, regardless of their valence. This seems part and parcel of how they function (i.e. that they enable us to deal efficiently with existing reality by ignoring subtle differences). This raises the question: is there, generally speaking, a clear point at which the harm done outweighs the benefits to the extent that we can qualify specific S&Ps as harmful in and of themselves?

It is difficult to see how stereotypes about large groups such as genders enable us to adequately deal with existing reality (they are just too crude to be helpful), and even when they do we might want to argue that the price to pay is too high. We could teach our children, for example, to mistrust all males when they get lost during a family outing, in a sense that might enable them to "adequately deal with existing reality," for it might reduce their chances of getting harmed on the rare occasions that they get lost. On the other hand, it might have very harmful consequences for their view of the world; for example, causing them to always be ill at ease around men they do not know. Stereotypes may also lead to what is called "stereotype threat"—a threat that has been shown to result in actual underperformance.[13] For example, the existing reality of women underperforming in mathematics seems to be the result of the stereotype that they are bad at it rather than a reflection of the stereotype's initial adequacy. (Research suggests that gender differences in mathematics performance in the general population are actually trivial; Hyde et al., 2008.) Stereotype threat also may cause underperformance due to racial or age stereotypes. African Americans, for example, are led to underperform in standardized tests when racial information is solicited prior to taking the test, thus making race salient to them (Steele and Aronson,

[13] See, for example, Stangor et al. (1998); Steele, James, and Barnett (2002); Thoman et al. (2008); and Fine (2010).

1995).[14] The elderly likewise underperform in memory tests when age is brought to the forefront (Chasteen et al., 2005).

When crude and undifferentiating stereotypes about the elderly, African Americans, and women become self-fulfilling prophecies,[15] they harm not only the individuals to whom the stereotype is misapplied, but the whole group that is targeted by the stereotype as well as society as a whole. It harms the individual female who excels in mathematics but is not recognized as such, but also, in a sense, women in general. Some women might have excelled at mathematics and enjoyed a subsequent career in that field but never explored that opportunity due to the S&P climate. It affects society as a whole—"harms" it in a sense—because it disables potentially capable individuals from excelling in an area they are not thought fit for. Note also that impaired self-esteem, stereotype threat, and self-fulfilling prophecies not only have an impact on the here and now, but also on the future. They all contribute to the maintenance of S&Ps and how ingrained they will remain in future cultures and societies.

In cases of impaired self-esteem, stereotype threat, and self-fulfilling prophecies it is clear and straightforward that harm is inflicted upon the individuals targeted with a negative stereotype. We have argued that harm is also done to many others beyond the targeted individuals. As a result it is in many cases not at all crystal clear and straightforward which people suffer or benefit from the S&Ps. Let us elaborate a bit more on this feature of S&Ps: the situation is complicated by the fact that both negatively and positively valenced stereotypes are at work.

Suppose, for example, that a female scientist gives a lecture and afterwards is approached by another scientist who compliments the male PhD student standing next to her for her performance by saying: "You can be proud with such talent in your group!" Although the other scientist acknowledges the excellent performance of the lecturer, she assumes the female scientist to be the PhD student and the male to be her professor. The female is automatically associated with being "less experienced" (a stereotype with negative valence) whereas the male is automatically associated with more experience and authority (a stereotype with positive valence). The example is framed in such a way to show that the stereotype is harmful to the female scientist; we can also frame it in a way to show that the male PhD student is also harmed. After all, because of a stereotype,

[14] The underperformance only occurred in so-called threat condition (in which Black participants in the test were told that the test was diagnostic of ability). In a so-called non-threat condition (in which they were told that the test was simply a problem solving not diagnostic of ability) there was no apparent risk of fulfilling the racial stereotype about their intellectual ability, and underperformance did not occur.

[15] See, for example, Snyder et al. (1977); Chen and Bargh (1997).

he is misjudged as being more experienced than he actually is, and people might subsequently judge him as "not very smart" because the standards by which they judge him are higher than appropriate.

Let us present another example of a stereotype working in two directions and causing harm of which we are not immediately aware. Male scientists are not standardly expected to contribute actively to household tasks and child rearing; few would think of complimenting a male scientist for the flourishing of his children, his household and/or how well he manages to combine his academic tasks with his domestic ones. In this sense, some men are dealt a short hand because of a stereotype. Moreover, it might even be the case that their careers suffer more from trying to combine those tasks than that of their female partners, because males are perceived as underperforming in comparison to their non-care-taking colleagues, while females are not. The point of these examples is to make clear that in an important sense, all people who do not fit the stereotypes that we tend to automatically apply to them are potentially harmed. Even though we tend to focus exclusively on the harm done to those targeted by negatively valenced S&Ps, the harm is not restricted to them.

Of course, there are other instances of negatively and positively valenced stereotypes that operate simultaneously, but in which the harm done is more clearly unevenly distributed and more harmful as in the case of stereotype threat. Research on the shooter bias paradigm, for example, shows that the stereotypes that Black people are aggressive and White people are not, both lead to misidentification of the objects Black/White males are carrying. In a simulated shooting task in a laboratory setting in which participants saw videotaped faces of men who were either holding a gun or a neutral object such as a wallet or a cell phone, more unarmed Black people were 'shot' because it was mistakenly inferred that the object in their hands was a gun. White people who were actually holding a gun were more often perceived as carrying a wallet and hence not shot at.[16] Clearly the shooter bias leads to disastrous consequences mainly for Black people, as is, unfortunately, illustrated outside the lab by numerous incidents of unarmed people of color[17] being shot in the public sphere of streets, parks, train stations, and so on, because of a misidentification of what they were actually holding in their hands.[18] We will elaborate on the role of biases in S&P behavior below, but first let us conclude this section.

[16] See, for example, Payne (2001); Correll et al. (2002); Greenwald et al. (2003).

[17] It is not only color of skin that might put you in danger. Islamic appearance and Islamic head-dress appear to have the same effect (Unkelbach et al., 2008).

[18] Cf. the Amadou Diallo case in New York City in 1999—see, for example, Kelly and Roedder (2008: 256)—or the Jean Charles de Menezes case in the aftermath of the London bombings of July 2005.

S&P behavior can lead to grave consequences. Even relatively neutral S&P behaviors that do not seem particularly harmful in isolation can result in considerable harm because of the accumulation of disadvantage that often goes hand-in-hand with it (Valian, 2005). We take it that there is considerable scientific and societal agreement on the potential harmful effects of certain S&Ps and that this is also the reason why there is considerable agreement that certain S&Ps should be discontinued. However, we should realize that S&Ps in general play an important role in our dealing with one another in a socially complex world. When adequate, they facilitate fast processing of enormous amounts of information. We should also realize that S&Ps are what we have called "collective constructs" and that the use of stereotypes often imply a simultaneous use of a "counterpart," an opposite or associated stereotype. As a consequence, it is not always clear-cut to whom the harm is done, though it might be very clear who suffers most (as in the shooter bias, and stereotype threat). Neither is it clear whether there is someone to blame for the harm done, even when it is clear who behaved in an S&P manner. That is, even though we might blame someone for behaving in an S&P manner, the harm that is done must be understood in the context of the wider societal setting.

Moreover, people who are harmed by S&P behavior—the victims, so to speak—might contribute to the continuation of the S&Ps; for example by "living up" to their stereotype. Think of, for example, a talented female mathematician who drops out to be a stay-at-home mom to "conform" to societal norms within her community. Doing so, she is a victim of S&Ps. At the same time, she contributes to a climate in which other talented women may feel pressure to conform to the stereotypical ways; in that sense, she is also a perpetrator. Hence, in a sense, not only are some perpetrators also victims, but some victims are also perpetrators. Also, some members of targeted groups share the S&Ps others hold about their in-group and engage in S&P behavior towards their own in-group members.[19] In addition, some members of targeted groups may display harmful S&P behavior towards people outside of their in-group. For example, women may manifest S&P behavior towards homosexual men. In both scenarios, the victims are also perpetrators. These observations are relevant because they might fuel the conviction that, with respect to harmful S&P behavior, everyone is guilty and everyone a victim, so why bother to think about changing things in this domain to begin with? If only to prevent such pessimistic conclusions, we believe

[19] Also, victims of S&Ps might be simultaneously or subsequently perpetrators because, for example, their self-esteem has been impaired because they have persistently been confronted with S&Ps.

it important to explain what is true about the claim and in what sense this still leaves ample room to try and change things.

In addition to the fact that victims and perpetrators are not clear cut categories where S&Ps are concerned, the nature and degree of harm itself also depends upon a complex mixture of the valence of S&Ps, the context in which they are applied and the individuals who are targeted by them. In the next section we turn to the question of what these complicating factors mean for our personal responsibility for disfavored S&Ps. The answer to this question, as we will see, is further complicated by the role of IBs in S&P behavior.

3 Responsibility and Implicit Bias

Let us first take a closer look at IBs, because there is no generally accepted definition of them and not all phenomena identified as IBs bear the same characteristics.[20] Within the context of S&Ps, "implicit bias" is often[21] understood to refer to (1) automatic association processes that function largely outside our awareness, (2) which affect how we "perceive, evaluate or interact with people from the groups that our biases 'target'"[22] on the basis of the particular S&P combination that is at work in a specific situation, and that (3) are understood to be prejudiced in the normative negative sense.[23] We will follow this usage, although strictly speaking the label 'implicit bias' refers to attitudes—normatively negative or not—that tend to steer us in a certain direction,[24] even though these attitudes might not be held explicitly or endorsed by the individual subject to them.[25]

[20] Cf. Holroyd and Sweetman, who argue in Chapter 1.3 of Volume 1 that differences between the various phenomena identified as instances of implicit bias are not to be ignored. These differences, they suggest, are likely to have considerable significance for the sorts of normative recommendations that are made regarding the mitigation of undesirable effects of implicit bias.

[21] For an excellent overview of empirical research and theoretical perspectives in the area of implicit intergroup bias, how they are represented in the mind, and how they are expressed in behavior, see Amodio and Mendoza (2010).

[22] Saul (2013).

[23] For example, Kelly and Roedder (2008); Kang and Lane (2010). One could call prejudiced behavior in the normative negative sense "discrimination." We will not do so, because we want to stay clear of any discussion about intent or criteria for disparate impact that are so prominently present in legal thinking about the notion of discrimination.

[24] Arguably, these attitudes *do* steer us, unless we fend off their impact—for example, when we anonymize in assessing CVs or grading papers to prevent that the attitudes might have an effect, or when we have succeeded in taking debiasing measures (e.g. implementation intentions).

[25] The literature on such implicit biases has grown enormously during the last few decades, and has attracted quite a lot of attention from the general public due to the fact that many findings show the limitations of our rationality; see, for example, Tversky and Kahneman (1986); Thaler and Sunstein (2008). We restrict our attention to S&P behavior, though our arguments might well serve

IBs typically involve associations between group-members and traits attributed to them. These biases may be understood as general automatic association processes. They function as a simplification; a shortcut for fast processing of information that can result in stereotypical and prejudiced behavior although we are largely unaware of their impact on our behavior. Research suggests that many people do have IBs and are prone to acting upon them.[26] When our behavior is affected by IBs, arguably we act unintentionally, in the sense that we do not consciously condone what we do.

IBs are based on culturally and historically influenced S&Ps that have become ingrained, for example, because of the 'standardized' way in which in a specific culture a trait is connected to a category (cf. Levinson, 2007).[27] For example, it is a culturally ingrained stereotype that women are naturally better at caring for children than men. This category (e.g. female) and these associative patterns (e.g. good at caring) have been so enduring that they, so to speak, obtain "a life of their own."[28] That is, they influence our behavior even though we are committed to egalitarianism on a conscious level. We here define egalitarianism as having consciously formed a commitment to a fair, respectful, and non-prejudiced treatment of participants in social interaction; that is, without giving any consideration to factors such as race, gender, age, and so on whenever these are not relevant to the situation at hand. Even when you are an egalitarian, IBs may influence you so that you show your baby to the one female in the company instead of the males standing next to you without thinking or even being aware of it.[29] The fact that IBs mainly function outside of our awareness is what makes the S&Ps so efficacious.

Is there any reason to believe that IBs actually play this role in our lives even in cases when we are committed to the just and fair treatment of one another? In

to clarify some issues with regard to our personal responsibility for what is called "faulty heuristics" or "cognitive impairments" as well.

[26] See, for example, Jost et al. (2009), who summarize recent empirical research revealing that nurses, doctors, police officers, employment recruiters, and many others exhibit implicit biases with respect to race, ethnicity, nationality, gender, social status, and other distinctions. See also the research on shooter bias and racially biased hiring practices that Michael Brownstein and Jennifer Saul point to in their general introduction to these volumes.

[27] There are also many implicit biases that are not culturally 'transmitted,' but instead derive from our "cognitive make-up" (see fn. 25).

[28] Although these social and cultural association patterns are what interest us in this chapter, we do not claim that there are no association patterns that are somehow hardwired and are unconnected to our specific culture; for example, the association of fire with heat, or a sudden loud noise with danger.

[29] This is an example that we devised to illustrate how it might work. It is not based on actual research, so other explanations for the behavior concerned might be possible.

their general introduction to these volumes, Michael Brownstein and Jennifer Saul point to abundant research on, for example, racially biased hiring practices or shooter bias that suggest there are indeed good reasons to believe this. In addition, they refer to a plethora of studies uncovering the pervasiveness of implicit biases against members of other stigmatized groups, such as women, gay people, and so on (cf. Jost et al., 2009). For reasons of space and to avoid unnecessary duplication, we refer to their editorial introduction for an explanation of the notion of implicit attitudes and the Implicit Association Test (IAT). For the purpose of our argument, it suffices to know that implicit attitudes are defined as "unidentified . . . traces of past experience that mediate favorable or unfavorable feeling, thought, or action toward social objects" (Greenwald and Banaji, 1995: 8). The IAT is the most popular method to measure implicit attitudes. It is understood to reveal our implicit attitudes, and in some cases may be more predictive of our behavior than self-reporting of our explicit attitudes. Notwithstanding the fact that the explicit attitudes of many people reflect strong and genuine egalitarian beliefs, research like that discussed in the editorial introduction suggests that most people hold non-egalitarian implicit attitudes. Many people are therefore prone to be affected by IBs and, at least at times, engage in S&P behavior. This leads to the troublesome conclusion that probably most people contribute in some way or other to the persistence of harmful S&Ps, regardless of the commitment to their discontinuation. Hence, when we are committed to the discontinuation of certain harmful S&Ps, part of that enterprise should also consist of the commitment to counteract IBs or prevent their efficacy.

We can, for example, take measures to discontinue such S&P behavior by prohibiting specific behavior, taking affirmative action and/or introducing certain procedures to prevent IBs influencing our practices. Anonymous grading is an example of the latter. Human beings are able—and often expected—to take responsibility for changing practices they disapprove of, taking into account their own inabilities. Using an alarm clock to wake up is a powerful and simple example of this. The fact that most of us do not wake on time, no matter how much we care for our jobs or other daytime commitments, is no reason to turn up late. By the same token, when IBs prevent us from discontinuing those S&Ps we are committed to discontinuing, we can be expected to take measures to prevent these IBs from influencing our behavior or take an interest in how they disable us from succeeding in what we are committed to do. We might even be expected to keep an open mind with regard to possible undesirable influences of IBs in areas in which there is, as yet, no empirical research. As Kelly and Roedder argue, when we have an "empirical hunch"[30] that it is more likely that we are affected by racial IBs than that we are not, this is a compelling epistemic reason to make some sort of compensatory adjustments; for example, correcting for the impact of IBs in grading the work of Black students. Those they call "savvy graders" might realize that Black students actually deserve a slightly better grade score; savvy graders might even use conscious rules such as bumping up borderline grades of Black students.[31] Given the ubiquitous presence of IBs,[32] it is reasonable to assume that we are affected by an array of them across the board.

To some, our argument for indirect personal responsibility might seem contrived. They may think it is obvious that we are directly responsible for our actions when they contribute to harm done. However, as we explained in the previous section, the harm done in the case of S&P behavior, even when it is clear-cut who suffers from it, does not always (and perhaps not even often) allow a simple identification of who is the victim or perpetrator. Take, for example, people who care about treating everyone in a fair and equal manner but still act in accordance with harmful S&P due to the influence of IB. They are perpetrators, but also, in a sense, victims, since they act in ways they explicitly reject. Hence, in some cases of harmful S&P behavior we might not be directly responsible or it might just not be clear whether we are. On top of this, as also explained in the previous section, S&Ps are part of our cognitive make-up in a way that does not make it easy for individual agents to abandon those S&Ps identified as harmful. The S&Ps that guide our judgments, actions, and evaluations are part of a complex mix of cognitive and affective components with different valences, depending for their application on the context. Moreover, we apply them often automatically and without preceding awareness. When IBs play the role they are now suspected to play, our individual control over harmful S&P behavior is even less than we thought. Nevertheless, we think we might still bear indirect personal responsibility on account of the collective nature of S&Ps. Since S&Ps are collectively construed and upheld, harmful ones can also only be collectively discontinued. As an individual we might succeed in not applying certain harmful S&Ps, but the real harm will only end once the harmful S&Ps themselves disappear. What we *can* do as individuals is contribute to the discontinuation of harmful S&Ps. Hence, if we are committed to the discontinuation of harmful S&Ps we can bear what we have called "indirect personal responsibility" for contributing to their discontinuation and for failing to do so. We can act in a praiseworthy fashion by behaving in ways that contribute to the discontinuation

[30] Kelly and Roedder (2008: 534).

[31] Kelly and Roedder (2008: 538, endnote 28).

[32] For example, Jost et al. (2009).

of harmful S&Ps. Conversely, we can continue to act in a blameworthy fashion by behaving in ways that fail to contribute to this discontinuation, or worse, by acting in ways that reinforce the harmful S&Ps.

We believe it is the collective aspect that makes the issue of our direct personal responsibility such a complicated affair. We also believe that this complicated nature is brought out by focusing on our indirect personal responsibility for harmful S&Ps, and that this talk of our indirect responsibility brings many more advantages worth spelling out. For example, people often feel they are treated unfairly when accused of showing S&P behavior, whilst the accusers might get angry with those whom they accuse of "adding insult to injury" by their outright denial of the characterization of their behavior as S&P behavior. We can do justice to both responses once we acknowledge that we have an indirect personal responsibility vis-à-vis harmful S&Ps. IBs, by their very nature, (generally) escape our individual awareness. As a result, when IBs cause us to engage in S&P behavior, it is understandable that we do not consciously *experience* ourselves as doing something wrong. We do not intend to display S&P behavior; moreover, we might feel indignant about the accusation because of our explicit egalitarian commitments. On the other hand, when you belong to the group that is harmed, it is understandable that you feel anger towards people who do not seem to take responsibility for the harm done. Often the S&P behavior—at least in the experience of those who feel harmed—is part of a pattern. Denial of the S&P character of certain behavior for them adds insult to injury because it suggests that they are not *victims* of a harmful practice. When you are familiar with the literature on IBs and are constantly met with a refusal to acknowledge responsibility, frustration is understandable as well. One might ask how others can fail to see and fail to care.

As we see it, emphasizing the fact that the enterprise of discontinuing harmful S&Ps is one to which each of us should contribute would be a more correct and also a much more fruitful approach to the issue. When we sidestep the issue of individual wrongdoing and blame, and focus instead on the reasons to take responsibility for harmful S&P behavior, as our argument for indirect responsibility does, different responses might be forthcoming. Instead of displaying a defensive attitude, people may be more inclined to look critically at whether S&Ps might have affected their doings (e.g. by them falling prey to IBs). Our argument for indirect responsibility suggests we take responsibility for the discontinuation of S&Ps on the basis of our role in social interactions and our egalitarian commitments (remember we use "egalitarian" here as shorthand for a fair and non-prejudiced treatment, i.e. rejection of harmful S&Ps). As such, it focuses on the future and the question of how, once we acknowledge the harm done, we can improve our practices. Clearly, one way to improve our actions is to scrutinize our own behavior critically for the effects of IBs. We come back to this in some detail in the next section.

Also, this indirect way of understanding our personal responsibility allows us to distinguish between degrees to which people can (be expected to) take responsibility and the occasions in which this is expected from them. The reason for this is that we do not all have a similar role in the social interactions that are affected by S&Ps and IBs. Suppose we are part of an informational and interactional environment that consistently and repeatedly expresses a commitment to gender equality; for example, the aforementioned example of academia. Tenured staff and students are among the constitutive members of the academic environment, although their role and impact in the institution differs because of the "power" they have to shape the institution, for example, by introducing and applying debiasing measures. In addition, their awareness of and experience with the phenomenon of IBs presumably differs, if only because of differences in age and in life in general.[33]

With respect to the collective effort to discontinue gendered stereotypes and prevent the efficacy of IBs that distort all kind of processes, it seems reasonable to expect those who are in a position of power to take more responsibility for change than those who are not. Those in power are the ones who invite authors or speakers for edited volumes or conferences, grade papers and examinations, evaluate CVs and research proposals for all sorts of purposes, and so on. Hence, they can be expected to have a stronger negative impact with respect to the continuation of harmful IBs and S&Ps and a stronger positive effect with respect to their discontinuation. Also, as more influential members of the academic environment, they are partly responsible for collectively upholding the S&Ps that prevent gender equality.

It makes sense to expect those in power to take more responsibility than, for example, students who are part of that same environment. The reason for this is not that those in power in comparison to students are better positioned or more able to counteract IBs, but that they have a different role in the collective of individuals responsible for the continuing influence of S&Ps. We think that talk about indirect responsibility better brings out the fact that different contributions can be expected at different times and occasions from different "players in the

[33] Nor are we all equally aware of the omnipresence of IBs and their impact on our behavior. We are not arguing that everyone should be aware of IBs to the same extent. Although the literature on IB research is skyrocketing, not everyone is familiar with that (yet). However, if one functions in a setting where equality (regarding gender, ethnicity, sexual preference, and so on) is part of the policy, pleas of a complete lack of familiarity with the phenomenon of IB would be unconvincing.

field," depending on their different roles within the collective.[34] At the same time, talk about indirect responsibility goes beyond a simple dichotomy between "those who are harmed" and the perpetrators of that harm. In our view, at least part of the tragedy of harmful S&Ps and IBs in a society with egalitarian aims and commitments is that there clearly is harm, but no clearly identifiable individual wrongdoers among those who subscribe to egalitarian principles. Except for explicit racists, sexists, or people who are in general not concerned with harm done, we are all, as it were, held captive by S&Ps we no longer endorse.

This brings us to a further benefit of our argument for indirect responsibility: it enables us to broaden our view on phenomena that contribute to or counteract our collective efforts at the discontinuation of S&P behavior. We sometimes take issue with people who behave in stereotypical ways or, on the other hand, applaud or feel proud of people who act contrary to stereotypes. On the basis of our argument for indirect responsibility, this is easy to understand. They contribute to the collective effort of discontinuing certain S&Ps or do exactly the opposite. Acting stereotypically, after all, contributes to consolidating the S&Ps in question that we want to get rid of, whereas acting contrary to stereotypes we want to get rid of actually contributes to the collective effort to discontinue them.

Let us conclude with a section on what we can do about behaving contrary to our consciously endorsed egalitarian commitments due to the efficacy of IBs.

4 Taking and Self-Ascribing of Personal Responsibility

If we are willing to take responsibility for and contribute to the collective effort of discontinuation of specific S&Ps, we might consider taking debiasing measures. But we might also go a step further; we could also scrutinize our behavior under the assumption that we have fallen victim to IBs. That is, we may examine our actions step by step in order to find out whether it was something we unwittingly did that caused the undesirable S&P outcome. We call this latter attitude "self-ascribing responsibility" to distinguish it from taking responsibility. Let us begin with ways in which we can take responsibility. As mentioned in the previous section, we can take responsibility in many ways. We can support legal and institutional debiasing measures, we can point out harmful S&P behavior, and we can try to avoid it ourselves. We can also attempt to diminish the undesirable

[34] See, for a nice taxonomy of the concept "responsibility" including that of role-responsibility, for example, Vincent (2011).

effects of IBs.[35] In the literature, various ways of debiasing are identified, some of them mutually reinforcing, which allow us to frustrate the process of automatic association and avoid the pitfalls of IBs.[36] Let us mention two of those, both of which focus on the idea of habits, and on how to break bad ones or form new ones.

Devine and colleagues started from the premise that IBs are like a habit that can be broken through a combination of awareness of IBs, a concern about its impact, and the application of strategies. They developed "habit-breaking intervention techniques" that they subsequently tested for efficacy, and found that the IAT scores of the intervention group were lower than those of the control group (Devine et al., 2012).[37] More importantly, and without precedent, they also found that the reduction in racial IBs lasted for at least two months after the intervention. According to the researchers, there was probably no single "magic bullet," but rather several components in combination resulted in the decline in racial IBs (Devine et al., 2012: 1277). It might well be that these habit-breaking techniques can also prove their worth in reducing S&P behavior in other categories (e.g. gender, age, and so on).

In addition to these habit-breaking techniques, another strategy looks promising: the habit-forming technique for debiasing known as *implementation intentions*. Implementation intentions are explicitly formulated if–then plans that supplement goal intentions; for example, not to be biased against women leaders (Webb et al., 2010). By forming implementation intentions, Webb and colleagues argue, we create a link between a specified opportunity (meeting a woman) and a response (treating her like a leader) that enables a quick identification of opportunities to initiate a new and "wanted" response to replace a stereotypical one. In forming implementation intentions, we therefore shed the effects of implicit associations so that we can "regain" self-control.[38]

In addition to taking debiasing measures, there is another way in which we can contribute to the collective effort of discontinuation of disavowed S&Ps: we could scrutinize our behavior under the assumption that we have fallen victim to IBs. More specifically, we may dissect our behavior step by step in an attempt to identify whether it was something we unwittingly did that caused the S&P outcome.

[35] For a comprehensive literature review, see Cornish and Jones (2013).

[36] See, for example, Kang and Banaji (2006); Jolls and Sunstein (2006).

[37] These techniques are: (1) replacing a specific stereotype by a non-stereotypical response, (2) imaging in detail positive counterstereotypic exemplars, (3) individuating out-group members by obtaining specific information about them, (4) taking the perspective of a member of an out-group, and (5) increasing opportunities for intergroup contact.

[38] See, for example, Henden (2008).

Let us illustrate the process of dissecting what we may have done by way of an example. Suppose a student approaches you to complain about her grade that, according to her, reflects your prejudicial assumptions about female students, rather than her personal mastery of the course material. You take a second look at her examination and find no reason to adjust your evaluation. When you check the results of all of your students, though, you notice that female students do score significantly lower than male ones. None of the individual examinations you re-evaluate give you any reason to adjust your original mark. Nevertheless the student's accusation, in combination with the bare facts (the unevenly distributed marks), will presumably unsettle you when you are familiar with how S&Ps and IBs work.

When you critically scrutinize your behavior as a result of the student's complaint, you might discover that a certain style of argument impresses you, and that this style happens to be employed more by male students than by female students. Subsequently, you might critically examine whether the "male" or "female" style really discloses a better grasp of the subject taught. If you find one style of argumentation is more effective, you can make explicit why you think that is the case and subsequently change your general instruction to students. You can, for example, be more explicit concerning what style of argument is expected and valued and for what reason. When students are made aware which style of argument will result in a lower grade and which one in a higher, they may adjust their style accordingly. On the other hand, if you come to the conclusion that the style you prefer does not disclose a better grasp of the subject, you might adjust your standards and evaluate the other style of argumentation more positively in the future. In both cases, it is very likely that the difference between the grades of female and male students will disappear. Note that this change is *not* due to a change of your attitude with regard to S&P behavior as such. Rather, it is due to your willingness to entertain two hypotheses: first, that it is something you do that may explain the gender-sensitive outcome, and second, that by pinpointing what that something is, you might be able to counteract this cause of the S&P outcome. To express it differently: by scrutinizing your judgments and doings you may uncover some effects of IBs that you were not previously aware of. We emphasize the distinction between "taking responsibility"—for example, by installing certain institutional practices to prevent biased outcomes—and what we want to call "self-ascription of responsibility"—i.e. seriously entertaining and investigating the possibility that it is something you have done, without being aware of it, that has produced an S&P outcome (Sie, 2008). A willingness to undertake a rigorous examination, we suggest, is an example of self-ascription of responsibility.

To be sure, the point of our example is not that we think specific reasoning styles exist for female and male students. What we want to point out is that there might be a variety of ways in which IBs affect us, and that scrutinizing our doings with that possibility in our minds might actually teach us something about ourselves and/or our practices.

5 Conclusion

In this chapter we explored the interrelations between stereotypes, prejudices, and implicit biases, and investigated when they are harmful and what this implies for our personal responsibility for them. We argued that in order to discontinue harmful stereotypes and prejudices, a collective effort is required. We can each contribute to this collective effort in a variety of ways. When we are committed to discontinuing harmful stereotypes and prejudices and we also believe that the role of implicit biases is what makes them so stubborn and difficult to change, we are also responsible for correcting or compensating for their influence. We looked at various debiasing mechanisms that enable us to frustrate the process of automatic association and decrease the impact of undesirable IBs on our behavior. Finally, we developed the notion of self-ascription of responsibility for disavowed S&Ps. We did so by exploring the example of a female student complaining about alleged biased grading of her work. The insight gained in this example is that we may learn things about ourselves by closely scrutinizing our behavior for the workings of IB that enable us to stay clear of the personal "pitfalls" that may trigger disavowed S&P behavior.

Acknowledgments

We would like to thank Jennifer Saul and Michael Brownstein for their very helpful suggestions and comments on earlier versions of this chapter, which, as a result, was improved considerably. We are also grateful to the two anonymous referees at Oxford University Press for their suggestions.

References

Allport, G. W. (1954). *The Nature of Prejudice*. Cambridge, MA: Addison-Wesley.

Amodio, D. M. and Devine, P. G. (2006). "Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior." *Journal of Personality and Social Psychology* 91(4): 652–61.

Amodio, D. M. and Mendoza, S. A. (2010). "Implicit intergroup bias: Cognitive, affective, and motivational underpinnings." In Gawronski, B. and Payne, B. K. (eds.), *Handbook of Implicit Social Cognition*. New York, NY: Guilford Press: 353–75.

Chasteen, A. L., Bhattacharyya, S., Horhota, M., Tam, R., and Hasher, L. (2005). "How feelings of stereotype threat influence older adults' memory performance." *Experimental Aging Research: An International Journal Devoted to the Scientific Study of the Aging Process* 31(3): 235–60.

Chen, M. and Bargh, J. A. (1997). "Nonconscious behavioral confirmation processes: The self-fulfilling consequences of automatic stereotype activation." *Journal of Experimental Psychology* 33(5): 541–60.

Cornish, T. C. and Jones, P. (2013). "Managing and reducing unconscious bias: What could work in the 'real world'? A Literature Review." Paper presented at the Bias Project Forum, Sheffield, 19–21 April.

Correll, J., Park, B., Judd, C. M., and Wittenbrink, B. (2002). "The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals." *Journal of Personality and Social Psychology* 83(6): 1314–29.

Devine, P., Forscher, P., Austin, A., and Cox, W. (2012). "Long term reduction in implicit bias: A prejudice habit-breaking intervention." *Journal of Experimental Social Psychology* 48(6): 1267–78.

Dovidio, J. F. and Gaertner, S. L. (1996). "Affirmative action, unintentional racial biases, and intergroup relations." *Journal of Social Issues* 52(4): 51–75.

Fine, C. (2006). "Is the emotional dog wagging its rational tail, or chasing it?" *Philosophical Explorations: An International Journal for the Philosophy of Mind and Action* 9(1): 83–98.

Fine, C. (2010). *Delusions of Gender: How Our Minds, Society, and Neurosexism Create Difference*. New York, NY: W. W. Norton.

Greenwald, A. G. and Banaji, M. R. (1995). "Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review* 102(1): 4–27.

Greenwald, A. G., Banaji, M. R., and Nosek, B. A. (2003). "Understanding and using the Implicit Association test: I. An improved scoring algorithm." *Journal of Personality and Social Psychology* 85(2): 197–216.

Henden, E. (2008). "What is self-control?" *Philosophical Psychology* 21(1): 69–90.

Holroyd, J. (2012). "Responsibility for implicit bias." *Journal of Social Philosophy* 43(3): 274–306.

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., and Williams, C. C. (2008). "Gender similarities characterize math performance." *Science* 231(5888): 494–95.

Jolls, C. and Sunstein, C. (2006). The law of implicit bias." *California Law Review* 94(4): 969–96.

Jost, J. T. and Banaji, M. R. (1994). "The role of stereotyping in system-justification and the production of false consciousness." *British Journal of Social Psychology* 33(1): 1–27.

Jost, J. T., Rudman, L. A., Blair, I. V., Carney, D. R., Dasgupta, N., Glaser, J., and Hardin, C. D. (2009). "The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore." *Research in Organizational Behavior* 29: 36–69.

Kang, J. and Banaji, M. (2006). "Fair measures: A behavioral realist revision of affirmative action." *California Law Review* 94(4): 1063–118.

Kang, J. and Lane, K. (2010). "Seeing through colorblindness: Implicit bias and the law." *UCLA Law Review* 58: 465–520.

Kelly, D. and Roedder, E. (2008). "Racial cognition and the ethics of implicit bias." *Philosophy* Compass 3(3): 522–40.

Kunda, Z. (1999). *Social Cognition: Making Sense of People*. Cambridge, MA: MIT Press.

Levinson, J. D. (2007). "Forgotten racial equality: Implicit bias, decision-making and misremembering." *Duke Law Journal* 57(2): 345–424.

Lippmann, W. (1922). *Public Opinion*. New York, NY: Free Press.

Macrae, C. N. and Bodenhausen, G. V. (2001). "Social cognition: Categorical person perception. *British Journal of Psychology* 92(1): 239–55.

McGarty, C., Yzerbyt, V. Y., and Spears, R. (2002). "Social, cultural, and cognitive factors in stereotype formation." In McGarty, C., Yzerbyt, V. Y., and Spears, R. (eds.), *Stereotypes as Explanations: The Formation of Meaningful Beliefs About Social Groups*. Cambridge: Cambridge University Press: 1–16.

Murray, G. R. and Schmitz, J. D. (2011). "Caveman politics: Evolutionary leadership preferences and physical stature." *Social Science Quarterly* 92(5): 1215–35.

Payne, B. K. (2001). "Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon." *Journal of Personality and Social Psychology* 81(2): 181–92.

Saul, J. (2013). "Implicit bias, stereotype threat, and women in philosophy." In Hutchinson, F. and Jenkins K. (eds.), *Women in Philosophy: What Needs to Change?* Oxford: Oxford University Press: 39–61.

Sie, M. (2008). "Intrapersonal ascriptions of responsibility." In Tze-Wan, K. (ed.), *Responsibility and Commitment. Eighteen Essays in Honor of Gerhold K. Becker*. Waldkirch: Edition Gorz: 47–58.

Snyder, M., Tanke, E. D., and Berscheid, E. (1977). "Social perception and interpersonal behavior: On the self-fulfilling nature of social stereotypes." *Journal of Personality and Social Psychology* 35(9): 656–66.

Stangor, C., Carr, C., and Kiang, L. (1998). "Activating stereotypes undermines task performance expectations." *Journal of Personality and Social Psychology* 75(5): 1191–7.

Steele, C. M. and Aronson, J. (1995). "Stereotype threat and the intellectual test performance of African Americans." *Journal of Personality and Social Psychology* 69(5): 797–811.

Steele, J., James, J. B., and Barnett, R. C. (2002). "Learning in a man's world: Examining the perceptions of undergraduate women in male-dominated academic areas. *Psychology of Women Quarterly* 26(1): 46–50.

Thaler, R. and Sunstein, C. R. (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven, CT: Yale University Press.

Thoman, D. B., White, P. H., Yamawaki, N., and Koishi, H. (2008). "Variations of gender: Math stereotype vulnerability to stereotype threat." *Sex Roles* 58(9–10): 702–12.

Tversky, A. and Kahneman, D. (1986). "Rational choice and the framing of decisions." *The Journal of Business* 59(4): 251–78.

Unkelbach, C., Forgas, J. P., and Denson, T. F. (2008). "The turban effect: The influence of muslim headgear and induced affect on aggressive responses in the shooter bias paradigm." *Journal of Experimental Social Psychology* 44(5): 1409–13.

Valian, V. (2005). “Beyond gender schemas: Improving the advancement of women in academia.” *Hypatia* 20(3): 198–213.

Vincent, N. A. (2011). “A structured taxonomy of responsibility concepts. In Vincent, N.A. et al. (eds.), *Moral Responsibility*, Library of Ethics and Applied Philosophy 27: 15–35.

Webb, T. L., Sheeran, P., and Pepper, J. (2010). “Gaining control over responses to Implicit Attitude Tests: Implementation intentions engender fast responses on attitude-incongruent trials.” *British Journal of Social Psychology* 51(1): 13–32.

1.5

Revisionism and Moral Responsibility for Implicit Attitudes

Luc Faucher

1 Introduction

Those acquainted with the literature about responsibility know that the notion that we are responsible agents has been questioned throughout philosophical history. More recently, psychologists and neuroscientists alike have also been arguing—due to research results—that responsible agency is under threat (Nahmias, 2006, 2010; Nelkin, 2005; Schlosser, 2013). In this chapter I will consider the possible influence of a subset of this research; namely, social psychology research on implicit attitudes, as it is related to our understanding of responsibility. The thesis I am defending is that certain revisions to our way of understanding responsibility, and the practices related to responsibility attribution, could be justified by results from research about implicit attitudes. My goal is modest; I am not trying to present an exhaustive list of revisions that might be induced by such work, nor am I trying to offer a unified view of such revisions. I am merely trying to illustrate—with a few examples—the type of revisions that could be induced by this research stream.

Prior to presenting these revisions (Section 5), I will elaborate on implicit attitudes, and the reasons why some psychologists see them as a threat to responsibility (Section 2). I then propose a taxonomy of forms of revisionism, and use this framework to qualify the revisionism I intend to defend (Section 3). Finally, I will discuss our intuitions about responsibility related to implicit attitudes, as well the ways philosophers have tried to capture them (Section 4).

2 Dual Attitudes

In psychology, the term "attitude" refers to a set of thoughts, feelings, or behavioral tendencies, with positive or negative valence, that we have towards a target object (person, social group, political party).[1] More precisely, by "attitudes" psychologists designate a set of heterogeneous mental states (stereotypes, emotional dispositions, or behavioral tendencies) that are (more or less firmly) associated with a target object. One topic that has become of crucial interest to social psychologists in recent years is that of "implicit attitude." One reason social psychologists have become interested in implicit attitudes is because many people nowadays consider the expression of certain prejudicial attitudes (for instance, in regard to people of other races) to be politically incorrect and, for this reason, may not want to express them explicitly for fear of being socially frowned upon. As a result of the latter phenomenon, the investigation of "explicit prejudice" is open to the possibility of self-presentation biases. It is also possible that people are unconsciously influenced by ambient cultural stereotypes and, while they express their sincere beliefs when they claim to be unprejudiced, in reality, they are in fact behaving with prejudice when interacting with individuals of other races (in these cases, we could talk of "introspective biases" because they would not be able to access and report their own biases).

To tap into implicit prejudicial attitudes (not directly accessible to the experimenter either due to the subject's social desirability concerns, or because these attitudes are unconscious) social psychologists have used "indirect" methods such as semantic priming, implicit association tests (IAT), affect misattribution procedure, and so on. (The term "indirect" is used to refer to the features of measurement procedures that "provide indicators of psychological attributes (e.g. attitudes) without having to ask participants to verbally report the desired information"; Payne and Gawronski, 2010: 4.) It has been shown that these implicit methods outperform explicit ones in predicting behaviors, choices, or judgments in socially sensitive domains (Greenwald et al., 2009; Pearson et al., 2009; Rudman and Ashmore, 2007).

According to some psychologists, implicit attitudes concerning a particular social group (scientist, obese, male, and so on) could be triggered by the mere perception of cues associated with membership in a group. For instance, John Bargh (1999) writes that "the mere perception of easily discernible group features (e.g. skin color, gender, and age-related characteristics) was sufficient [...] to cause the activation of the stereotype associated with the group, which then was

[1] My use of "attitude" is following Hardin and Banaji (2013: 15).

shown to influence judgments of a group member in an unintended fashion, outside of a perceiver's awareness" (363). More recently, he wrote that in certain cases "the mere presence of certain events and people automatically activates our representations of them, and concomitantly, all of the internal information (goals, knowledge, affect) stored in those representations that is relevant to responding back" (Bargh and Morsella, 2008: 76).

Why should we care about automatically activated attitudes? Their existence would not be of much concern if they had no impact on our behavior. Yet, unfortunately, as Ferguson and Bargh observe, research in social psychology has begun to demonstrate that "complex behavior is also automatically shaped and guided by the knowledge that is incidentally activated during perception" (2004: 33). Implicit attitudes have been shown to negatively influence helping behavior, hiring, legal, or medical decisions, interracial interaction, cooperation, political decisions, shooting decisions, and so forth (Pearson et al., 2009; Payne, 2006). These attitudes, then, are causally involved in the production of morally problematic behaviors.

One important question becomes determining how much of our lives are controlled by these attitudes. While some, like Keith Payne (2006; see also Pearson et al., 2009), seem to think that the influence of attitudes is restricted to particular situations (for instance, when cognitive or attentive resources are depleted or weakened), others such as Bargh counter that "[...] most of a person's everyday life is determined not by their conscious intentions and deliberate choices but by mental processes that are put into motion by features of the environment and that operate outside conscious awareness and guidance ..." (Bargh and Chartrand, 1999: 462). As we can infer from the last portion of the quote, Bargh is skeptical that we can control these attitudes. On one hand, conditions necessary to control them (for instance, the awareness that stereotypes are operating, motivation to act on them, and knowledge of how to neutralize them) are rarely met. On the other hand, even when in place, "stereotypic judgments and behavior can nonetheless occur" (Bargh, 1999: 371). So, according to Bargh "the evidence to date concerning people's realistic chances of controlling their automatically activated stereotypes weighs heavily on the negative side" (378).[2]

If the situation is as Bargh describes it, one would be justified in believing that the domain of actions for which we are responsible suddenly becomes very

[2] For a similar, and more recent, pessimistic assessment of our capacity to control implicit attitudes, see Hardin and Banaji (2013: 16). In Section 5.3 I will present some reasons to be more optimistic.

narrow—much more so than most of us imagine. Indeed, if a large part of our everyday actions essentially escape our control and are completed despite us, one could be tempted to say that we are not their source, and therefore we are not responsible for the bulk of our daily actions. This is what Bargh has deemed the "tremendously depressing implication." As he puts it: "How can anyone be held responsible, legally or otherwise, for discriminatory or prejudicial behavior when psychological science had shown such effects to occur unintentionally?" (Bargh, 1999: 363).[3] For reasons such as this, certain researchers consider the work of social psychologists like Bargh to be a threat to responsibility. If this is indeed the case, the threat is not directed at our concept of responsibility as such, but at its presumed extension. If what these social psychologists claim is true, we would be under an illusion concerning the range of actions we are responsible of. We would be responsible for much less than we think—if responsible for anything at all.

However, could we not accept the idea that part of our life is guided by implicit attitudes and reject the tremendously depressing implication? In other words, could we not hold ourselves responsible for actions caused by our implicit attitudes? It seems, prima facie, that to be held responsible for actions caused by our implicit attitudes would force us to go against some of our intuitions concerning responsibility—but would it? Could our concept of responsibility allow us to be responsible for actions that we did not voluntarily perform? It seems that to answer this question we need more clarification in regard to the concept(s) that we are employing. This is what I will do in Section 4. Prior to this—because I wish to propose a revisionist approach in this domain—I will explain what I mean by "revisionism."

3 Revisionism

My intention is to argue that research about implicit attitudes could induce some revision to the way we think about responsibility, and I would like to be clear about what I mean by "revision," and what kind of revisions I intend to propose. My understanding of the question is largely due to Manuel Vargas' framework (see 2005 for a statement of his position). Ergo, in the following, I will present his position in broad strokes.

[3] Wellman (2007) agrees with this judgment: "If bias is ultimately a function of biology and neurology, human actors do not control it. Consequently, they cannot be held accountable for discriminatory behavior [. . .] it might provide the ground for an effective defense against allegations of discrimination . . . " (50).

Vargas claims that when philosophers are interested in the question of responsibility, they are typically interested in one or more of the following aspects: 1) dispositions or attitudes associated with responsibility (for instance, reactive attitudes, such as indignation, anger, or guilt); 2) practices associated with responsibility (for instance, blame or punishment); and, 3) beliefs or concepts concerning responsibility (for instance, beliefs about the normative conditions that have to be met for attributions of responsibility).

In regard to these elements, philosophers typically ask three kinds of question. First, *metaphysical* questions: "What is the nature of responsibility?" Second, *descriptive* questions: "What are the conditions that we think need to be met to apply our concept of responsibility?" Third, *normative* questions: "How should we think about responsibility?" For non-revisionists, there is a sort of "pre-established harmony" (Vargas, 2005: 411) between what we think about responsibility and what we should think about it, given what we know about responsibility itself or about the world. A revisionist does not share this belief and asserts that there is a difference between the descriptive and the normative projects, such that what the second prescribes differs from what the first describes. In other words, and as Vargas puts it: a theory is revisionist if it "prescribes a revision, relative to a diagnostic account of commonsense thinking, in our responsibility-characteristic practices, attitudes, beliefs, or our conception of these things" (2005: 421).

In Vargas' view, the strength of revisionism can vary, so we may speak of weak, moderate, or strong forms of revisionism. Revisionism will be labeled "weak" if it is not our concepts, practices, or attitudes that are revised, but rather our understanding of what they are (after we discover that we were mistaken in what we thought were the folks' concepts, practices or attitudes). In contrast, revisionism will be labeled "strong" if it proposes the elimination of our concepts, practices or/and attitudes, for instance, because they are too metaphysically demanding, or because it is psychologically impossible to meet what they demand. "Moderate" revisionism—when applied for instance to our concept of responsibility—"is the idea that the folk concept of responsibility is inadequate until it has been modified in some way . . . it does not involve straightforward elimination of the concept, practices, or attitudes characteristic of responsibility. Rather it amounts to a "pruning" of that element" (2005: 409). It could, for instance, consist of the elimination of an element in our concept of responsibility which has been judged to be superfluous, like the concept of self-determination.

Not only strength, but also the range of revisionism can vary. For instance, what Vargas calls "sophisticated revisionism" (2005: 412) allows for the revision of one element of what is of concern when we think about responsibility (for example, we might want to drop certain practices, like punishing), without modifying the other elements (like our concept of responsibility or our reactive attitudes, like moral indignation). Moreover, the sophisticated revisionist might have a different attitude towards each element of responsibility: for instance, they might want to throw out our concept of responsibility, but keep attitudes intact, and modify certain practices (for instance, keeping blame but discarding punishment). A sophisticated revisionist can also propose varied changes within a category. Vargas (2005) provides the example of the revision of the concept of responsibility: "Consider," he says, "the various possible targets for purely conceptual revisionism: epistemic conditions, or the kinds of things a responsible agent must know; the freedom condition, or the kind of freedom a responsible agent must have; the ultimacy condition, or some notion that the considered act or state of affairs is ultimately up to the agent; various considerations about rationality; and presumably, a capacity for consciousness. In principle, any of these things are open to deflation or elimination, in any combination" (2005: 413).

What are the constraints that one should respect when proposing a revision? Vargas claims there are two constraints: the revision should be 1) plausible and 2) justified normatively. Concerning plausibility, he writes that "a revision that requires something that is not psychologically possible or socially implausible is likely to be and perhaps ought to be rejected. Similarly, revisionisms that rely on highly speculative or largely implausible accounts of agency will fare worse than revisionisms that do not rely on dubious pictures of agency" (413). He later writes that in the case of a weak revision, the justification for revision will often originate from empirical studies (given that these have a higher degree of epistemological authority than our own beliefs in regards to a concept): "to the extent that... empirical accounts alter our everyday understanding of, say the practices or attitudes characteristic of responsibility, such theories will count as weakly revisionist theories of responsibility" (414). In other words, empirical studies could demonstrate that individuals' concepts of responsibility are different from the concept we "as philosophers" are using. In terms of normative justification, the task could consist (for example) of showing that despite a major revision of our concept of responsibility (or its total elimination), we are still justified—on the grounds of evolutionary and utilitarian reasons—in retaining our attitudes (such as guilt) and practices (such as blame). We could also make reference to other aspects of ethical theory (for instance, considerations of justice or equality) to justify a revision.

In what follows, I will propose that work on implicit attitudes contains the seeds of revisions to the way in which we understand various aspects of responsibility. For this reason, I therefore argue for some form of "sophisticated revisionism." The revisions I propose are weak or moderate revisions, and are both psychologically plausible and justified by empirical research.

4 Folk and Philosophical Conceptions

Let us return to the question of responsibility for implicit attitudes. Keith Payne nicely summarizes our problem: "The problem with split-second decisions [the ones that are prompted by implicit attitudes] is that they seem to make themselves [. . .] Should I consider those decisions my decision, if they differ from my intentions? Who is responsible?" (Payne, 2006: 287). We could add: "Am I responsible?" "Am I to be blamed for them?" "Should I feel guilty about them?" Note that in what follows I will generally not focus on or debate our responsibility for *having* implicit attitudes. Rather, following Brownstein (in preparation), I will focus on our responsibility for the *behavioral expression* of our implicit attitudes (BEIA).

Understandably (given that the idea that our behavior might be guided by implicit attitudes is fairly recent), not much has been done until recently to probe our folk intuitions concerning responsibility in BEIA cases. The only empirical study on the folk conception of responsibility and implicit attitudes is Cameron et al. (2010). In this study, Cameron and his colleagues are taking advantage of a debate in the implicit attitudes literature concerning the character of implicit bias. Indeed, while many take implicit attitudes to be unconscious (for instance, Greenwald and Banaji, 1995), some have questioned this assumption and have defended the idea that while the influence of implicit attitudes on behavior might be automatic or outside one's control, implicit attitudes, as such, might be consciously accessible, for instance, in the form of "gut feelings" (Gawronski et al., 2007; or Nier, 2005; I will say a bit more about that in Section 5.2). In Cameron et al.'s (2010) study, participants had to read either vignettes where an agent (explicitly opposed to racial discrimination) discriminates based on implicit attitudes that were outside the agent's control but that were either unconscious or conscious (the authors present the contrast as between "sub-conscious dislike" and "gut feeling dislike"), or a vignette describing identical behavior without mentioning implicit attitudes. Participants then rated the agent's degree of responsibility on a scale of 1 to 5. What these researchers found is that "[o]nly discrimination resulting from unconscious bias was excused, suggesting that *conscious awareness matters for judgments of moral responsibility*" (2010: 278; my emphasis). Simply put, only unconsciousness seems to exculpate. Participants consider someone who is conscious of their implicit attitudes but unable to control them almost as responsible for their actions as someone whose acts are not under the influence of implicit attitudes. So for participants, the fact that the discrimination is out of the control of the agent does not excuse as such. They discounted moral responsibility only for discrimination due to unconscious bias. Why is that so? Why does consciousness play such a role in moral responsibility attribution? This is what Levy (2008: 213) calls the "hard problem" of moral philosophy. Since it is part of a philosopher's job to elucidate the concepts we use in daily life, let's see what philosophers have to say about the reasons why unconsciousness might be an exculpating factor in the aforementioned cases.

At present, there are only a handful of philosophers who have been preoccupied with the question of responsibility for BEIA. Among those, at least one seems to be (and only to a certain extent) in agreement with folk conception as depicted by Cameron et al.: Neil Levy (2014; but see also his 2008 and 2012, and Levy and Bayne, 2004; for a similar view, Wigley, 2007).[4] For him, "moral responsible agency is . . . conscious agency" (2008: 214). Levy considers himself as advocating a "volitionist theory" of responsibility because in his view, conscious decisions or voluntary control are the crucial elements for the attribution of responsibility.

According to Levy (2008), one cannot be held responsible for a BEIA. As he writes:

> In absence of consciousness we are at the mercy of automatic responses we are not responsible for acquiring, that we may consciously reject and which we may even have worked hard to eradicate. It is therefore unfair to hold us responsible for actions which reflect such responses when we cannot control them; thus, consciousness is necessary for moral responsibility [. . .] When our actions are the direct product of our conscious choices [. . .] we are responsible for them. (219)

Levy seems to suggest two reasons why we cannot be held responsible for our actions when they are caused by our implicit attitudes. The first reason has to do with the fact that implicit attitudes are not conscious; the second one with the fact that they are not under our control.

The first reason is that because we are unaware of our attitudes,[5] we cannot give them our approbation or oppose them, and we cannot endorse or reject

[4] Other philosophers with a different take on responsibility of implicit attitudes are Holroyd (2012) and Kelly and Roedder (2008). See additional chapters in this volume.

[5] As I said earlier, in Section 5.2 I will revisit the assumption that implicit attitudes are under the radar of consciousness. The second reason for rejecting responsibility for BEIA that I attribute to Levy does not require making this assumption.

them.[6] The idea that we are only responsible for actions if they proceed from volitions or attitudes that we endorse, or that we have chosen after deliberation, is an idea defended in literature by the likes of Frankfurt (1971) or Watson (1975). As Levy (2008) puts it:

> Actions are deeply attributable to agents only when consciousness plays a substantial role in their production ... Conscious deliberation is properly reflective of the entire person, including her consciously endorsed values. (217 and 220; see also his 2012: 243)

In other words, an action is deeply attributable to the agent when it results from conscious deliberation.[7] It is only in this condition that the action is the expression of the agent's "evaluative stance" (2012: 244), where an evaluative stance is an agent's "global perspective on what matters, normatively" (Levy, 2012: 255). In other words, for an agent appropriately to be held responsible, a belief or an attitude that will be the basis of action must be able to be evaluated according to reasons or ends that the agent endorsed. This is obviously not the case with implicit attitudes, which are not "personally available"[8] to the agent and for this reason cannot be evaluated in light of agent-endorsed values or ends.

The second reason is that, because we are not conscious of these attitudes, we cannot exercise the proper kind of control required for moral responsibility over BEIA.[9] Since we cannot control these behaviors, we cannot be held responsible for them. This second reason is in fact a classical condition of responsibility (mentioned by Aristotle in Book Three of his *Nicomachean Ethics*, the condition of control). An agent is said to be responsible for an action if and only if he or she is not forced or constrained in this action. Levy's condition is a version of the control condition because he seems to posit that consciousness is necessary for control. But the fact that one is conscious of something does not ensure that one can control it. (For instance, I might be conscious of some of my deeply ingrained habits and not be able to control them; or someone suffering from alien hand

[6] As Levy (2012) remarks: "When we act unconsciously, attitudes like our belief that the action is wrong may fail to be activated, so only some smaller proportion of our moral attitudes is expressed by the action" (254).

[7] I will say more about what Levy means by "deep attributability" in Section 5.2.

[8] According to Levy (2012), "[i]nformation is personally available when it is so readily available [to the agent] that it requires little effort to retrieve and it is poised to guide behaviour" (246–7). One might argue that the condition invoked by Levy holds only as long as the agent is not made aware of the existence of his implicit attitudes. Levy could reply that this kind of awareness is not the kind of consciousness needed for responsibility (for instance, because one is not aware precisely when the implicit attitudes are influencing behaviour and when they are not). Alternatively, he could also provide another reason why the agent should not be held responsible. The second reason offered by Levy hereafter might be interpreted as such a reason.

[9] As Levy and Bayne (2004, 2013) put it: "consciousness enables us to inhibit and veto our initial impulses to act."

syndrome might be conscious that his hand is doing something, but not able to control its movement directly.) Therefore, one could argue that the problem with implicit attitudes seems to come more from their presumed automaticity than from the fact that they are unconscious. Levy recognizes this in his "Consciousness, Implicit Attitudes and Moral Responsibility," where he claims that the problem with implicit attitudes might not be that they are unconscious, but rather that they fail to integrate with our "personal level concerns in the manner required for [. . .] moral responsibility"(ms., 9).[10] Thus, the problem with implicit attitudes seems to be that they are not constrained by our deliberate decision not to be prejudiced. (Levy calls them "judgment-insensitive" because "they stubbornly refuse to fall into line with our consciously endorsed beliefs"; 2012: 256.) In this, he seems to go against what Cameron and colleagues attribute to folks, i.e. that unconsciousness is the main exculpating factor and that automaticity does not exculpate much. But given the role that automaticity played in history as an exculpating factor (for instance, in forensic psychiatry; see Reznek, 1997), it is not clear that it is not part of folks' conception either.

5 Revisions

In this section I will make three revisionist proposals. Once again, to be clear, I do not intend to propose a global revision of our concept of responsibility, nor do I intend to present a unified set of proposals. Rather, I will suggest some revisions to our current concepts (as well as to certain philosophical conceptions that are avowedly in continuity with our folk concepts) that research on implicit biases inspires. Will these revisions inspire more profound, radical revisions? Will they clear the road for the elimination or replacement of our concept? While it is possible, at present I cannot and will not answer these questions. My wish is to convince skeptics that reflection on implicit attitudes should lead to substantial revisions in our concept of responsibility.

In Sections 5.1–5.3 I will propose three revisions, related to

1) the role of conscious knowledge or conscious intention in the attribution of responsibility;
2) the real self theory, as proposed by Frankfurt and Watson (also by Levy in the context of our discussion);
3) the concept of control used in discussions of responsibility.

[10] Elsewhere, Levy (Levy and Bayne, 2004) writes that "[a]gents who act automatically are not responsible for their actions because they cannot exert the right kind of control over their action, where the right kind of control is control that manifests character in a deep sense" (214).

Once again, the revisions that I propose are rather weak ones (indeed, using Vargas' scale, they are either weak or moderate, but by no means strong)—that is, revisions in our way of understanding our own concepts or modifications to our concepts that do not lead to their elimination or replacement.

5.1 Conscious knowledge and conscious intentions

As stated earlier, the fact of being conscious of the moral demands of a particular situation seems to be a necessary condition for the attribution of responsibility. For this reason, the fact that agents do not know that they have implicit attitudes and that these attitudes are acted on despite them, would exculpate the agents of the BEIA. Levy captures and makes explicit these folk intuitions when he emphasizes the fact that knowledge of the attitude is a necessary condition for the attribution of responsibility.

The problem is that there might be individual or group differences in folk intuitions. In other words, it is possible that the idea that it is necessary to be conscious of something in order to be responsible for it might not be shared by all. Indeed, as Cameron et al. (2010: 278) remark in a footnote, while it is true that for Caucasian participants lack of consciousness exculpates, African Americans seem to take a different view. Though it is premature to conclude anything from these results (due to a small population of African Americans in the referenced study; i.e. a total of seventeen over the ninety-five participants), there seems to be a trend in the literature demonstrating that a group that is the subject of discrimination is more likely to attribute responsibility to dominant group members,[11] even when there is no conscious or explicit intention to harm. For victims of discrimination, the intentions of the perpetrator are not the focus of interest; what counts is the fact that some harm has been done. For instance, law professor Barbara Flagg (1993) writes that "White people tend to view [conscious] intent as an essential element of racial harm; non-whites do not," (968) or Janet Chan (2011), in a work on racial profiling, remarks that "[f]or victims of racial profiling, the intention of the policing agent is not an issue [. . .]" (75). Here is how I think we should understand these claims. Whites will oftentimes deny any wrongdoing because they acted on the basis of good, conscious, intentions (or at least, not on the basis of an intention to harm). But from a Black person's perspective, good intentions do not make any difference: the action done is offensive and/or unjust. And not being able to recognize or realize the character of one's actions (because they were made in good faith, they cannot be that bad!), or trying to evade responsibility by arguing that one had good intentions (or no

[11] See fn. 13 for some works documenting this trend.

bad intentions), is even more offensive.[12] Also offensive is trying to blame the victims for denouncing discrimination, as often happens (and calling them racists for paying attention to race as a factor in discrimination while Whites believe this is not a factor any more, for instance; see Mills, 2007: 28). Responsibility has to be taken for actions and who else can take responsibility if not those who performed the offensive or unjust action? (After all, it is their mental states, their internal negative representations of the member of a racial group that caused the action.) Finally, I understand the reaction of certain Whites who urge that responsibility should be taken for their unconscious biases as a reply to this call for responsibility.[13]

This makes me suspect that there might be a difference between discriminating and discriminated-against groups in the way they attribute responsibility. In short, for individuals who are the victims of racism on a regular basis (or who are part of an oppressed group), the presence of an conscious intention to harm might not be required for an action to be morally reprehensible and the presence of good intentions is not enough to exempt or exculpate (so someone who has good intentions and who knows that he has a bias, or a dislike, would not be exempted from responsibility, nor would the one who do not knows about his bias or dislike). As for Whites who perform racial discrimination, intentions, because more accessible, might be more central. As Pearson et al. (2009) explain:

> Whites' perceptions about how they are behaving or how they are perceived by others are based more on their explicit attitudes and overt behaviors [. . .] and less on their implicit attitudes or less deliberative behaviors. In contrast, the perspective of Black interaction partners in these interracial interactions allows them to attend to both spontaneous (e.g. nonverbal) and deliberative (e.g. verbal) behaviors of Whites . . . (9)

Pronin (2007) calls the *general* tendency to give increased weight to more accessible introspectible data in the evaluation of our behaviors, and less weight

[12] Choosing not to take responsibility for discrimination's harm when no explicit intention is involved is a problem, and not wanting to change the "intent doctrine" can also be perceived as blameworthy. Flagg (1993), whose paper is about discrimination law, remarks that "[. . .] retaining the intent requirement in the face of its demonstrated failure to effectuate substantive racial justice is indicative of a complacency concerning, or even a commitment to, the racial status quo that can only be enjoyed by those who are its beneficiaries—by white people" (969).

[13] For instance, Shannon Sullivan (2007) writes that "[c]haracterizing white privilege as increasingly operating through unconscious habits, in other words, is not a way to let white people off the hook for their non-reflective racism or their implicit acceptance of benefits of white privilege. It instead is a claim that notions of responsibility have to be rethought to encompass unconscious habits" (233), or Bob Samuels (2010), in his blog, argues that " . . . we cannot base personal responsibility on what people intend in a conscious way; rather, people must be held accountable for their unconscious associations." <http://www.huffingtonpost.com/bob-samuels/unconscious-racism-at-the_b_491817.html>.

to publicly observable external data (such as behavior), the "bias blind spot." She asserts that "[p]eople generally believe that they are immune to group-based biases. They claim freedom from racial bias [. . .] even in circumstances where they have shown these biases [. . .]" (38).

What this suggests is that we not only suffer from blindness in regard to certain situations (i.e. we do not seem sensitive to the moral demands of certain situations), but we also suffer from blindness in regard to our blindness (a kind of "moral blindness denial"). We would not recognize the traces of our prejudices in our own behavior, nor in the behavior of those who belong to our group.

As such, Pronin describes a general tendency towards blindness, i.e. a tendency to prefer a certain type of data when it comes time to evaluate our own behavior. What I wish to suggest is that the fact of being the subject of discrimination may well redirect attention from what agents are thinking, to what they are doing (accentuating the tendency of those who are the subject of discrimination to use external rather than internal data).[14] If this is the case, an asymmetry between points of view might be induced by the fact that we are committing an injustice (or a moral harm), or that we are being treated unjustly (or harmed). In other words, we are on different sides of the fence; our perspectives and interests are different, and we are attentive to different things.

This leads me to my first suggestion of revision. This revision is inspired by a recent discussion of moral responsibility by Knobe and Doris (2010; for a similar proposal, though in a different form, see Glasgow, this volume). In their discussion, Knobe and Doris argue that recent work in experimental philosophy shows that our attribution of moral responsibility depends on and changes as a function of certain contextual variables. They call this idea "variantism," as opposed to

[14] One possible explanation of this effect might be the fact that discrimination makes salient the fact that we do not belong to a certain group, and therefore see more clearly the actions or effects of the group's actions (for instance, in limiting our tendency to put ourselves in the shoes of this group's members to interpret their behaviors). See Bruneau et al. (2012) for a description of one possible mechanism behind this effect. An alternative explanation might be related to what is called in the social psychology literature the "asymmetry hypothesis." According to the asymmetry hypothesis, "... discriminatory acts perpetrated by the strong against the weak will be seen as more biased than similar acts perpetrated by the weak against the strong. In the context of interracial encounters, this hypothesis suggests that White people, who historically comprise the dominant racial group in America, will be viewed as more racist when treating members of minority ethnic groups in an unpleasant manner than would members of a minority ethnic group when treating White people in an unpleasant manner" (Marino et al., 2010: 641). In cases such as these, expectations play a role in shaping the way we categorize action. Other contextual factors, identification with one's group (Sellers and Shelton, 2003), enduring conditions of domination (Coutant, 2006), or endorsement of individual mobility beliefs by victims (or the meritocratic worldview; Kaiser and Major, 2006), might all have an impact on the sensitivity to stigmatization and perception of discrimination.

"invariantism." They argue that invariantism has dominated philosophical discussions up to now, and consists of the idea "that people should apply the *same* criteria in *all* of their moral responsibility judgments. In other words, it is supposed to be possible to come up with a single basic set of criteria that can account for all moral responsibility judgments in all cases." However, they observe that

> [i]t seems that people do not make moral responsibility judgments by applying invariant principles. Instead, it appears that people tend to apply quite different criteria in different kinds of cases. Thus if one wants to understand why people make the judgments they do, it is no use looking for a single basic set of criteria that fits all people's ordinary judgments. A more promising approach would be to look at how and why people may adopt different criteria in different cases, depending on the way an issue is framed, whether an agent is a friend or a stranger, and so on . . . (322)

Ergo, I would like to make the following proposal: one of the variables that has an impact on our judgment of responsibility for BEIA is the fact of having been or being (perhaps on a regular basis) the victims of these BEIA. I would call this idea "variantism in function of discrimination" (VFD). Indeed, I propose that we should test VFD further using the Cameron et al. paradigm, putting participants in a condition where they are the objects of discrimination, or making salient the fact that they belong to a group which has been historically discriminated against[15] by the group making hiring decisions. Depending on the results of the experiment,[16] I propose that we could face a "weak" revision of our concept of responsibility. Contrary to what we might have thought, we might well be variantists. Our judgments might be unstable; they might be affected by a contextual variable—i.e. the fact of being or having been the object of discrimination on one side and being the "discriminator" or (as in Cameron et al.'s original experiment) being an observer of a discrimination by an in-group member against a out-group member, on the other side.[17]

[15] Because embracing one's ethnic or racial identity plays an important role in perception of discrimination (Marino et al., 2010: 658), we would expect the effect to be found only for groups with whom the participants identify.

[16] Results of the experiments on the asymmetry hypothesis make me think that the kind of experiment I am proposing will support VFD, though, Coutant's (2006) work suggests that it will work mainly for groups who have been discriminated against stably through time, not for punctual episode of discrimination.

[17] Here, results of experiments on the so-called *intergroup attribution bias* indicate that VFD might well be supported. According to intergroup attribution bias, "[. . .] positive outcomes associated with an in-group actor and negative outcomes associated with an out-group actor would be attributed to causes rated as internal, stable, uncontrollable by others, and global. In contrast, in-group-negative and out-group-positive outcomes would be attributed to causes rated as external, unstable, controllable by others, and specific" (Islam and Hewstone, 1993: 936). This bias is

The literature reviewed here seems to demonstrate that victims of moral harms see things differently from perpetrators. We can interpret victims' judgments in two ways—only one of which supports variantism. The first consists of saying that victims judge those who discriminate (while disavowing discrimination) as hypocrites. The second consists of saying that, for those who are the subject of discrimination, the presence of conscious intentions is not essential. According to this second interpretation, we would be justified in attributing to those who suffer from discrimination a theory of responsibility different from the one that puts consciousness at the center of responsibility, when the subjects of discrimination judge those who discriminate against them. Here, again, it seems that there are two possibilities. First, it is possible that in certain cases, we tend to switch from a responsibility theory in terms of "attributability," to responsibility in terms of "accountability." Watson (1996) introduced this distinction, and puts it thusly: "Holding people responsible is not just a matter of the relation of an individual to her behavior; it also involves a social setting in which we demand (require) certain conduct from one another and respond adversely to one another's failures to comply with these demands" (262). So we might not have conscious intention to harm, but others can hold us responsible for it, in light of moral demands that we should meet but have not yet met. This possibility is explored by Zheng (this volume) and Brownstein (ms.), so I will not comment further on it. The second possibility is that individuals experiencing discrimination are adopting a position similar to that defended by George Sher in a series of papers (2006, 2008) and in his work *Who Knew?* (2009). I shall expand on this a little in what follows.

Sher claims that in our daily life we hold people responsible for more than they are aware of, or for more than they consciously intend to do: I can hold you responsible for forgetting my birthday, for not being sensitive to what I am saying to you, or for having forgotten the dog in a hot car—without at any time thinking that you were consciously doing so, or that you had the intent of doing so.

Sher thinks—akin to what Pronin proposed—that there is an asymmetry in attributions of responsibility, and that it is not at all clear that first-person and third-person attributions are being made on the same basis. As Sher (2009) puts it, when judging the actions of an agent:

at work when members of perpetrator groups explain their ancestor's misdeeds: "When explaining historical crimes, descendants of perpetrator groups often use biased attributions. They perceive historical crimes as caused by situational factors and as unstable, and the perpetrator group ashighly variable [...] Another exonerating strategy among descendants of perpetrator groups is to blame historical victims for their fate" (Vollhardt and Bilewiz, 2013: 4).

> From our retrospective and external vantage point, his conscious beliefs have no particular priority over his physical makeup or unconscious attitudes or traits. Hence, as long as blaming and holding people responsible are reactions that we have to them [the people] from a perspective that does not coincide with their own, the way they must view their choice situations when they deliberate will provide us with no obvious reason *not* to base our blame or attributions of responsibility on facts about them and their choice situations of which they were not aware. (61)

What is the basis of our judgment of responsibility? Sher proposes a neo-Humean theory of responsibility that aims to capture the essence of our judgments of responsibility for actions that we do not consciously intend. It is neo-Humean, as Sher believes that the link between an agent and his action is the character of the agent, from which the action flows—but his notion of character is extremely thin. As he explains in response to a comment by Levy: "On my account, what renders an unwitting wrongdoer responsible is not that his failure to realize that he is acting wrongly expresses a cavalier attitude toward such wrong acts, but only that that failure of recognition can be attributed to some combination of his causally effective states that may or may not include such an attitude" (2008: 225).[18] To be clear, Sher is not simply collapsing Watson's "responsibility as attributability" and "responsibility as accountability" into one form of responsibility akin to accountability. As he says: "I want to consider the possibility that when an agent should, but does not, recognize that he is acting wrongly or foolishly, what connects him to the act's wrongness or foolishness *is not just his failure to live up to whatever standard requires that those in his position recognize such acts as wrong or foolish* [which is what responsibility as accountability seems to come down to], but is rather the whole collection of attitudes, dispositions, and traits whose interaction causes him not to recognize this" (2009: 87, my emphasi). This set of attitudes, dispositions, and traits, conscious or unconscious, is part of what Sher considers a "person" (2008: 223). According to Sher, our intuitions about control are driven by a theory about agents understood in terms of "authorship or origination," so that "to say that a certain feature of what an agent did was within his control is to say, at a minimum, that his performing an act with that feature can be traced back to *him* as opposed to some aspect of the situation or circumstances" (224). This would be

[18] Elsewhere he writes that the failure of an agent to recognize "his act's wrongness need not to have anything to do with [his] basic values or commitments but may instead be fine-grained cognitive or inferential patterns that he neither endorses or recognized" (2006: 299) or "unlike the non-culpably ignorant agent, who simply lacks any information that would support the conclusion that he is acting wrongly, the culpably ignorant does have the necessary information, but is prevented by some aspect of his character or belief-system from putting the pieces together" (2009: 21).

consistent with the current belief that one of the first steps in improving our conduct and our character is taking responsibility for actions that resulted in failures to live up to certain moral standards, even if we did not recognize them as morally wrong when we performed them. This would also explain why some of us might feel guilty when we learn that some of our actions have been prompted by our implicit attitudes. (See Glasgow, this volume, for an illustration of such a case.)

Sher (and Pronin) suggests a general position according to which there is an asymmetry between the first- and third-person judgments; my proposal is more circumscribed. I suggest that a variable that affects the judgment of responsibility is the fact of being the subject of discrimination. In these conditions, victims do not take into account the conscious or avowed intentions of those who discriminate, instead taking into account their actions. I have suggested that it is possible to interpret the victim's stance in two ways: either (1) they consider those who discriminate (despite conscious disavowals of discrimination) as hypocrites or as being of bad faith;[19] or (2) do not necessarily presuppose that those who discriminate are conscious of what they do, but judge that their actions are a reflection of the person they are (while construing "person" quite liberally). As per the first interpretation, victims would establish responsibility on the basis of presumed unavowed or hidden intentions. According to the second interpretation, victims would judge the character of those who discriminate; they would hold those who discriminate responsible, if they think that discrimination emanates or originates from mental states inherent to them. If I were correct that victims shift to the latter theory of responsibility, this would militate in favor of a particularly moderate form of revisionism. Firstly, it would argue in favor of a form of variantism similar to that proposed by Knobe and Doris (as long as the victims would hold a different theory of responsibility in a non-oppressive context, or shift back to a theory of responsibility that places conscious intentions at its center when in a dominant position for long enough). Secondly, the theory employed by victims might also be revisionist in regard to our usual concept of responsibility. If, according to our concept of responsibility, responsibility is predicated on conscious intentions or conscious knowledge, then Sher's view of responsibility is a revision of our concept.[20]

[19] Davidio et al. (2002) suggest that this might be the case: "Blacks may attribute the behaviors of Whites to explicit rather than implicit prejudice, thereby assuming that Whites' negative behavior is intentional [. . .]" (67).

[20] Sher explicitly considers this possibility in his (2009: 151).

5.2 *"Real self" theories*

In Section 5.1 I defended the idea that criteria applied to individuals to determine their responsibility might be "variable," and that for this reason it could support the idea that individuals have radically different theories of responsibility. I proposed that in certain contexts, theories based on conscious intentions cannot explain our responsibility judgments. Still, such theories sometimes do the work—that is, they sometimes capture our intuitions. I address in this section potential revisions induced by work on implicit attitudes related to the place given to consciousness in responsibility.

In "Restoring control" (2008), Levy comments on types of cases of actions that we are interested in, that we could be considered responsible for because this responsibility requires "deep attributability." As we saw in the previous section, actions are deeply attributable when "consciousness plays a substantial role in their production" (217). Generally, this is played out through conscious deliberation, such that "conscious deliberation is properly reflective of the entire person, including her consciously endorsed values" (220). Susan Wolf (1993) has labeled theories of that kind "real self" theories. According to this type of theory (defended by Dworkin, 1976; Frankfurt, 1971; Watson, 1975; among others), an agent cannot be held responsible if their actions do not accord with or do not flow from consciously adopted ends or values; in other words, if they are not expressive of what the agent stands for.

A classic objection to this view (see Arpaly, 2003; Arpaly and Schroeder, 1999; Thalberg, 1978) consists of asking why we should privilege a conscious deliberation in the establishment of what constitutes my real self, or, as Thalberg elaborates about such theories:

> Both Frankfurt and Dworkin assume that when you ascend to the second level,[21] you discover the real person and what she and he really wants. I shall pack my misgivings in [a challenge]: Why grant that a second-order attitude must always be more genuinely his, more representative of what he genuinely wants, than those you run into at ground level? (1978: 219–20; he makes similar arguments against Watson's non-hierarchical version as well)

It is thus possible to question my seriousness in terms of what I claim to consider important (or value), and therefore, what I consider mine and what I reject as not-mine. In the context of our discussion, we can ask ourselves if the dissociation between our explicit and implicit attitudes might not be sometimes simply

[21] The "second-level" Thalberg mockingly alludes to is what Frankfurt calls "second order volition."

the result of agent confusion (the extent of this confusion should be determined empirically) concerning what is asked of them in an experimental context. Perhaps when they are asked about their explicit attitudes, they are in fact referring to what they believe is desirable, what is socially expected rather than to what they really endorse—their true motivations. In order to know if we are responsible for our BEIA, it seems we should be able to answer the following questions: How am I to determine that an attitude is really mine, as opposed to not-mine? How am I to know that a motivation is really mine, that it reflects what I really endorse? As Fisher (2012b) noticed recently: "The problem is that the hierarchical account does not have the resources to distinguish between an individual's real self (or real self for the purposes of practical reasoning) and the individual's "ego ideal"—that is, what he wants to be or perhaps aspires to be,"[22] and because of this, "[o]ne at least needs to supplement the hierarchical apparatus with elements that provide the resources to distinguish the real self from the ego-ideal..." (137). In what follows, I would like to present two lines of research that demonstrate that methods developed in the domain of implicit attitudes could provide such resources.

The first line of research concerns the possibility that dissociations between implicit and explicit attitudes were created by self-image concerns. The idea is that it is possible that the source of variation between these two kinds of attitudes originates from the fact that in explicit attitude tests, we can control our responses and we may want to paint a picture of ourselves that conforms to social expectations. Thus, these studies do not permit the discovery of our true motivations or real attitudes, but rather motivations or attitudes that are socially acceptable. Nier (2005) tries to work around this concern. In his study, he uses a strategy used previously by Sigall and Page (1971) in one of their studies. The idea is the following. Participants are told that the experimenter has in possession a means to detect what their true attitudes are (in Nier's case, he told certain participants that his implicit attitude test would reveal their true attitudes). This procedure (called the "bogus pipeline") would have the advantage of minimizing the strength of social desirability as it relates to what participants are saying (with their lies detected, participants are better off being truthful). In this condition:

[22] Frankfurt seems to accept Fisher's diagnostic as he wrote: "Nor is it likely to be readily apparent whether a decision which a person intends to be wholehearted is actually so. We do not know our hearts well enough to be confident whether our intention that nothing should interfere with a decision we make is one we ourselves want carried out when... we come to understand more completely what carrying out would require us to do or to sacrifice doing" (1998: 175–6).

> [T]he results indicated that when participants believed their "true attitudes" were being accurately assessed, there was a significant relationship between an implicit measure of racial attitudes (the IAT) and an explicit measures (the MRS [Modern Racism Scale]) [. . .] Thus the results suggest that as the motivation to report explicit attitudes that are consistent with implicit attitudes increases, the implicit–explicit relationship strengthens due to changes in self-reported explicit attitudes [. . .] (2005: 48–9)

Later in his paper, Nier argues that his results contradict the idea that the origin of the dissociation between implicit and explicit attitudes comes from the fact that they are two different types of representations, due to two different types of processes. As far as access to consciousness is concerned, implicit and explicit attitudes could, to a certain extent, be similar.

This line of work suggests that the implicit attitudes of a subject are not really alien to them, that they know that they have them.[23] Being conscious of these attitudes, one might consider that they should try to eradicate or at least to neutralize them. One might think that someone will eradicate or neutralize these attitudes if they are truly motivated to do so. How can we be sure that one is truly motivated to do so, that one is motivated *not* to endorse implicit or explicit attitudes?

Here again, work on implicit attitudes can assist us, and provide a key tool to answer this question. To understand how this tool works, we have to understand a distinction used by psychologists between *internal* and *external* motivation. Amodio (2008) describes this distinction as such:

> Our prior research had recently shown that participants' implicit racial biases—which at that time were assumed to be inevitable and immutable—varied as a function of their motivations to respond without prejudice. That is White Americans report that they respond without prejudice towards Black people for two independent reasons—in order to meet their personal, internal standards [*internal reasons*], and to avoid negative reactions from others who may disapprove prejudice [*external reasons*] . . . Devine et al. found that people who were motivated only by internal reasons [. . .] consistently exhibited lower levels of implicit bias on reaction-time measures than participants reporting each of the other motivational profiles . . . (12)

In this context, it becomes important to be able to measure internal motivations. How are we to know that one is really motivated internally? The question is relevant, as it is always possible—because of social pressure again—to say that

[23] For similar conclusion, see also work by Uhlmann and Nosek (2012) and Payne, Burkley, and Stokes (2008). Madva (in preparation) proposes that the type of consciousness we have of our implicit biases might be similar to Ned Block's (1995) "phenomenal consciousness;" that is, we are conscious of some aspects of our implicit biases (the affective element for instance) but we do not pay explicit attention to them.

one is motivated by internal reasons while indeed being motivated by external reasons. It is thus to internal reasons that we need to have access, if we want to measure genuine motivation.

How can one access internal motivations while simultaneously ensuring that they are not simply a *façade*?[24] One way consists of measuring "implicit motivations," supposing that they reflect the subject's *real* motivations (because they are not influenced by the preoccupation of looking good in the eyes of others).[25] How could we measure these implicit motivations? In "Implicit motivation to control prejudice" (2008), Glaser and Knowles propose a method to provide such a measurement. They assert that implicit motivation to control prejudices (IMCP) is composed of two elements: a negative attitude towards prejudice (NAP), and a belief that oneself is prejudiced (BOP). The standard implicit attitude test can measure each component. For the NAP component, the test consists of measuring the strength of the association between words like "tolerance" or "prejudice" and "good" or "bad" (or equivalents). For the BOP component, the test consists of measuring the strength of association between words like "tolerant" or "biased" and words like "me" or "stranger" (or equivalents). What these measures reveal is both surprising and interesting:

> Those high in BOP and high in NAP are the only ones who exhibited a non-positive (in fact, a slightly negative) relation between race weapon stereotypes [RWS] and the shooter bias.[26] Those low in NAP and high on BOP show the strongest positive relation

[24] This question is not only of rhetorical interest as some studies have shown that doubt about the fact that positive actions are really motivated by internal motivations not to be prejudiced are widespread in certain minorities groups and that these doubts are responsible for negative reactions to positive actions (Major et al., 2013).

[25] I am talking in terms of "*real* self" and "*façade*" because this is how the debate is framed (it is also because of concerns of self-presentational biases that implicit measures of self-concept have been introduced; see Schnabel et al., 2008), but as I will try to show at the end of this section, the literature I review might force us to revise this way of talking about the self. I will posit, for the sake of discussion, that parts of our "real self" are the motivations which we "deeply" identify with (by deeply here, I mean not only at the explicit level, but also at the implicit level), but also which guide our actions (as shown by the work I review, implicit motivations seem to play a role in guiding actions in contexts where explicit motivations typically fail). As Schnabel and Asendorpf (2010) observe: "Implicit procedures have the potential to provide access to aspects of the self that are inaccessible or only partially accessible by conscious introspection but that may have an essential impact on how one thinks, feels and behaves" (408).

[26] Correll et al. (2002) and Payne (2001; see Payne, 2006, for a summary of that research) have shown that when people stereotypically associated a racial group with danger (or in Glaser and Knowles' experiment when a racial group is stereotypically associated with weapon, this is what they call the "race weapon stereotype"), they have a lower response threshold when having to judge that members of that racial group are carrying a weapons. In standard experimental paradigms, participants are presented pictures of Black and White targets holding either a weapon or benign object like a soda can or a cell phone. Their task is to identify as quickly and accurately as possible what type of object the target is holding. They have then to shoot if the target is holding a weapon

> between RWS and shooter bias, perhaps reflecting the implicit motivation to use stereotypes (i.e. prejudice is ok [low NAP], I am prejudiced [high BOP], and I am going to use it!). (19)

In other words, what Glaser and Knowles's study demonstrates is that those with a high IMCP are not displaying the terrible and devastating effects characteristic of those who harbor implicit racist attitudes. In these individuals, the presence of a stereotype by which Blacks are more likely to have guns than Whites does not lead to faster firing on Blacks than Whites (or to making increased false positive identifications).

In short, the studies I have discussed here reveal two things. Firstly, that our implicit attitudes may not be "under the radar" of consciousness; they may not be estranged bodies in our psyche—we know that we have them, and that they are ours. This fact does not command a revision of our concept of responsibility, but rather a revision of a particular understanding of implicit attitudes. Moreover, if these attitudes are conscious, the fact that we do not reject them wholeheartedly could well make us responsible for them—at least according to criteria from "real self" theoreticians and also folk understanding. This may be the interpretation we should attribute to participants of the Cameron et al. study: if one is conscious of one's attitudes, but does nothing to neutralize their influence, one then proves that one does not reject these attitudes wholeheartedly. Yet how are we to know that we are really rejecting these attitudes? After all, it is only when we possess a real (internal) motivation to remain uninfluenced by these attitudes that we can neutralize them. Herein lies the second lesson learned from these studies: psychology provides a means of deciding which motivations an agent deeply identifies with. The revision that I propose is thus the following. When we seek to discover an individual's internal motivations (these are the kind of motivations that an agent accepts on the basis of reasons that are deemed important to him, so they are *his* in the "real self" theorists sense), it is important not to focus only on what he says, but also on his implicit motivations. Explicit motivations may be expressions of what an agent aspires to, but reference to implicit motivations might be a better way to know how deeply ingrained these motivations are in his psyche. It might be tempting to conclude from this that implicit motivations

and do not shoot if the target is holding a benign object. What Payne has shown is that people who hold stereotypes about the danger of a racial group respond more quickly, and more mistakenly, to pictures of members of that group; that is, they are quicker to recognize weapon in their hands and make more mistakes thinking that they are holding a weapon when they have a benign object in their hands. This is what Correll and his colleagues have termed the "shooter bias" (Payne refers to it as the "weapon bias").

reflect the "real self," but that would be too quick.[27] A rather more modest conclusion would be to say that when it comes time to evaluate our motivations, we should be aware of the fact that motivations comes in different kinds and different depths (internal/external, explicit/implicit), and that social psychology is helpful both in revealing them and in proposing ways to assess them.[28] This revision is thus not a revision of our concept of responsibility, but rather a revision of the way we decide when one is responsible, and it is a revision that supports those who believe that the "real self" should be involved in responsibility attribution (but not everybody does so; see, for instance, Fisher, 2012a). Because it does not question the role of motivation in responsibility, but introduces a new way to conceptualize motivation, we could describe this revision as "moderate."

5.3 The notion of control

Since Aristotle's classical conception of responsibility, control has been considered one of the necessary conditions for responsible action. For some, control necessitates consciousness, such that the condition of control could not be met in the absence of consciousness. If such is the case, it is quite unlikely that we are responsible for our actions when they flow from our implicit attitudes, for as Hardin and Banaji (2013) put it:

> [...] research shows that it is nearly impossible to consciously correct for effects of implicit prejudice [...] To do so, one must be in the unlikely circumstance of having at once (a) knowledge that implicit prejudice is operating, (b) both the motivation and cognitive capacity to control it and perhaps more unlikely of all, (c) precise knowledge of the magnitude and direction of the correction needed... (16)

If such is the case, it would comfort those who believe we are not to be held responsible for our actions when they are produced by implicit biases (and those who, like Bargh, draw a "tremendously depressing implication" from that). However, this might not be necessarily be the case. The literature addressing the question of control of BEIA suggests that the three conditions stipulated by Hardin and Banaji do not need to be met for control to be present. In fact, there

[27] What psychologists call implicit motivations (as well as explicit ones) are parts of an agent "self concept." It is how he conceives of himself. This conception can be erroneous or not very accurate (I might think of myself as being a party person, while in fact I am not). But importantly, as has been shown, these representations are not epiphenomenal; they have an impact on behavior. For this reason, they are, in a sense, part of one's real self (though that may be more in the rather liberal sense of "self" that Sher was arguing for; see Section 5.1).

[28] Though some more questions will need to be investigated concerning the correlation between explicit and implicit measures of this self-concept, as well as about the stability and malleability of implicit motivations.

might be a greater chance that control works if some of these conditions are *not* met. As Pearson et al. (2009: 16) note:

> [...] whereas conscious efforts to avoid stereotyping may often fail or even exacerbate bias because individuals lack insight into the processes that promote and regulate it, passive implicit goals to not stereotype may succeed by co-opting the very psychological mechanisms that sustain it, replacing stereotypic associations with egalitarian or atypical associations when perceiving or interacting with members of other racial and ethnic groups. (16)

I would like to present two ways to control implicit attitudes that do not necessitate consciousness of the fact that they are operating, nor motivation (at least at one level) to apply them, or even knowledge of the magnitude of control that we need to exercise. In short, these methods do not demand that we know the mechanisms or methods employed to counteract them. In this way, they are "indirect" (Holroyd, 2012: 286). As for the first one, Holroyd observes that it requires "intentional undertaking, so [it] will only be relevant to cases in which individuals know they are biased and seek to mitigate this" (286). The second method, though tested in a lab, might well work in certain conditions in the wild and would not demand the recognition of the presence of implicit attitudes, but at least the recognition of a failure to live according to norms that we try to live by.

In the literature about control, two forms of control are usually distinguished: top-down control, and bottom-up control. These forms of control are distinguished in terms of requisites. Top-down control is a control that the agent manifests when they decide consciously, in a particular situation, not to let their attitudes dictate or influence their behavior. Bottom-up control is a control that does not depend on the recognition of the activation of attitudes or intentions to inhibit one's behavior. In what follows, I will present two forms of bottom-up control: the "implementation of intentions" and the "priming of egalitarian goals." I will then pose some questions about the links between these kinds of control and responsibility.

Stewart and Payne (2008; see Mendoza et al., 2010, for a slightly different version of the same experiment) use the implementation intentions method as a mean to reduce automatic stereotyping. This method consists in asking subject to repeat sentences in the "if... then" form in which a cue is linked to a goal. It is postulated that in this condition, the goal gets linked to a cue that indicates at which moment the goal should be active. It has been shown that this method works better than forming abstract and general intentions (like "I should avoid being biased").

In one of their experiments, Stewart and Payne asked to subjects to repeat sentences such as "Whenever I see a Black face on the screen, I think the word 'safe'." Comparing results obtained using this method with results obtained by using sentences that do not emphasize the reduction of attitude (such as "Whenever I see a Black face, I think 'accurate' [or 'quick']"), the authors have shown that only subjects in the first condition were not demonstrating effects characteristic of the "shooter bias" (see fn. 15), but rather of the opposite effect—that is, they took additional time to recognize a gun in the hands of an African American man than in the hands of a White man.

The second method—the priming of egalitarian goals—is proposed by Moskovitz and Li (2011). In their paper, the authors examined "if egalitarian goals that are *not chronically held* lead to control of stereotype activation by triggering operations that include the inhibition of stereotypes" (105). Inspired by ideas developed in the literature on goal selection, according to which the contemplation of a failure in a domain activates a goal, Moskowitz and Li asked participants to reflect on a past experience of failure in these two different domains: in egalitarian action, or with respect to traditions. In each case, participants were invited to think about a time when they failed either to fairly treat an individual, or failed to respect a tradition. This experiment revealed that where participants remembered a time when they failed to treat an individual fairly, control of the stereotypes follows. Once again, "[s]tereotype control is evidenced by this effect disappearing so that stereotypic words no longer are responded more quickly, and instead, due to spreading inhibition, are responded more slowly following faces of Black men" (107).

In light of the two research studies mentioned, I propose the following revision: unconscious processes can be subject to significant kinds of control. This is what Suhler and Churchland (2009) proposed recently when they wrote: "... nonconscious processes can support a robust form of control and, by extension [...] consciousness is not a necessary condition for control" (341). Moreover, it would appear that the mechanisms in charge of these forms of control are "sensitive to moral reasons." They can be put to work by the consideration of moral reasons.

This kind of control raises one question. Is this form of control sufficient to establish responsibility? I do not see why not. In the case of the first kind of control, the fact is that we do have control, but not direct control; rather we have what Holroyd calls "long-range control." In the second case, it is a bit more complicated. Priming of egalitarian goals demands that we really have some goals and that these goals get "activated" by a perceived or remembered failure to achieve them. What is necessary then is to have these goals activated; control then follows without any conscious involvement. A similar kind of unconscious control seems to be displayed by people with internal motivation (Section 5.2). Remember: in their case implicit attitudes do not have the same impact on behavior than they have on those who are externally motivated. According to

Holroyd, "these considerations support the claim that individuals can indirectly and unintentionally control the manifestation of biases even when they are unaware of the possibility of influence" (288). It is not at all settled whether people would be willing to judge as responsible those who display the kind of indirect and unconscious control (that is, control that does not require the participation of conscious practical deliberation) we presented in this section, but as my account is revisionist, I see no problem in going against people's intuitions (and some philosophers are willing to go along with me; for instance, Fisher and Ravizza, 1998: 86).

If we accept the idea—as presented in Section 5.2—that we could have implicit motivations and that these could be our real motivations, and if we accept Payne and Moskowitz and Li's results, it seems that it is indeed possible to trigger these motivations that control our behavior, without any conscious involvement on our part. Furthermore, difficulties evoked by Hardin and Banaji do not appear to force us to accept the tremendously depressing implication drawn by Bargh. If this implication was the consequence of our concept of responsibility, where control can only be conscious and voluntary, then what we have just presented forces a revision of the condition of control: control does not have to be direct; it does not even have to be conscious. This revision is a weak one because it does not question the condition of control, but rather questions its interpretation (which I believe is its classical interpretation).

6 Conclusion

I argued that certain revisions to our way of understanding responsibility could be justified by recent work on implicit attitudes. Inspired by some distinctions introduced by Vargas, I distinguished various types of revisionism. I have been arguing here for a sophisticated form of revisionism (either of the weak or moderate strength). Thus, I did not argue that we should get rid of our concept of responsibility. I hope I have convinced my readers that revisions need to be undertaken, yet if I have failed to do so I hope that at least I have at least demonstrated that the existence of implicit attitudes poses serious difficulties for our concept of responsibility, and that work on implicit attitudes can be used to create finer distinctions related to notions involved in our reflection on responsibility.

Acknowledgments

I wish to thank those who offered helpful comments where I presented a version of this chapter, including the University of Latvia in Riga, the University of Western Ontario, Sheffield University, Université de Montréal, and the Centrer of Neuroethics of the Institut de Recherches Cliniques de Montréal. I would like to offer special thanks to the editors of this volume and to the evaluators of my papers, and also to Sarah Arnaud, Aude Bandini, Lawrence Blum, Amanda Leigh Cox, Stéphane Lemaire, Edouard Machery, Ron Mallon, Christine Tappolet, and Paul Taylor for the comments on previous versions of this chapter. This work has been made possible by a grant from the SSHRC.

References

Amodio, D. M. (2008). "The social neuroscience of intergroup relations." *European Review of Social Psychology* 19: 1–54.

Arpaly, N. (2003). *Unprincipled Virtue: An Inquiry into Moral Agency*. New York, NY: Oxford University Press.

Arpaly, N, and Schroeder, T. (1999). "Praise, blame and the whole self." *Philosophical Studies* 93(2): 161–88.

Bargh, J. A. (1999). "The cognitive monster: The case against controllability of automatic stereotype effects." In Chaiken, S. and Trope, Y. (eds.), *Dual Process Theories in Social Psychology*. New York, NY: Guilford Press: 361–82.

Bargh, J. A. and Chartrand, T. L. (1999). "The unbearable automaticity of being." *American Psychologist* 54: 462–79.

Bargh, J. A. and Morsella, E. (2008). "The unconscious mind." *Perspectives in Psychological Science* 3: 73–9.

Block, N. (1995). "A confusion about consciousness." *Brain and Behavioral Sciences* 18(2): 227–47.

Brownstein (in preparation). "Attributionism and responsibility for implicit biases."

Bruneau, E. G., Dufour, N., and Saxe, R. (2012). "Love, hate and indifference: Behavioral and neural responses in Arabs, Israelis and South Americans to each others' pain and suffering." *Philosophical Transactions of the Royal Society: Biology* 367: 717–30.

Cameron, C. D., Payne, K., and Knobe, J. (2010). "Do theories of implicit race bias change moral judgments?" *Social Justice Research* 23: 272–89.

Chan, J. (2011). "Racial profiling and police subculture." *Canadian Journal of Criminology and Criminal Justice* 53(1): 75–8.

Correll, J., Park, B., Judd, C. M., and Witterbrink, B. (2002). "The police officer's dilemma: Using ethnicity to disambiguate potentially threatning indiviuals." *Journal of Personality and Social Psychology* 83: 1314–29.

Coutant, D. (2006). "The effect of a power-imbalanced situation of the cognitive processing of low-power group members." *Group Dynamics: Theory Research, and Practice* 10(1): 71–83.

Davidio, J., Kawakami, K., and Gaertner, S. (2002). "Implicit and explicit prejudice and interracial interaction." *Journal of Personality and Social Psychology* 82(1): 62–8.

Dworkin, R. (1976). "Autonomy and behavior control." *Hasting Center Reports* 6(1): 23–8.

Ferguson, M. J. and Bargh, J. (2004). "How social perception can automatically influence behavior." *Trends in Cognitive Sciences* 8(1): 33–9.

Fisher, J. M. (2012a). "Responsibility and autonomy: The problem of mission creep." *Philosophical Issues* 22(1): 165–84.

Fisher, J. M. (2012b). "Semicompatibilism and its rivals." *Journal of Ethics* 16: 117–43.

Fisher, J. M. and Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Flagg, B. (1993). "Was blind, but now I see: White race consciousness and the requirement of discriminatory intent." *Michigan Law Review* 91(5): 953–1017.

Frankfurt, H. (1971). "Freedom of the will and the concept of a person." *Journal of Philosophy* 68: 5–20.

Gawronski, B., LeBel, E., and Peters, K. (2007). "What do implicit measures tell us? Scrutinizing the validity of three common assumptions." *Perspectives on Psychological Science* 2: 181–93.

Glasgow (this volume). "Alienation and responsibility."

Glaser, J. and Knowles, E. D. (2008). "Implicit motivation to control prejudice." *Journal of Experimental Social Psychology* 44: 164–72.

Greenwald, A. G. and Banaji, M. R. (1995). "Implicit social cognition: Attitudes, self-esteem, and stereotypes." *Psychological Review* 102: 1, 4–27.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E., and Banaji, M. R. (2009). "Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity." *Journal of Personality and Social Psychology* 97(1): 17–41.

Hardin, C. D. and Banaji, M. (2013). "The nature of implicit prejudice: Implications for personal and public policy." In Shapir, E. (ed.), *The Behavioral Foundations of Public Policy*. Princeton, NJ: Princeton University Press.

Holroyd, J. (2012). "Responsibility for implicit bias." *Journal of Social Philosophy* 43(3): 274–306.

Islam, M. R. and Hewstone, M. (1993). "Intergroup attributions and affective consequences in majority and minority groups." *Journal of Personality and Social Psychology* 64(6): 936–50.

Kaiser, C. and Major, B. (2006). "A social psychological perspective on perceiving and reporting discrimination." *Law and Social Inquiry* 31(4): 801–30.

Kelly, D. and Roedder, E. (2008). "Racial cognition and the ethics of implicit bias." *Philosophy Compass* 3(3): 522–40.

Knobe, J. and Doris, J. (2010). "Responsibility." In Doris, J. and the Moral Psychology Research Group (eds.), *The Moral Psychology Handbook*. New York, NY: Oxford University Press: 321–54.

Levy, N. (2008). "Restoring control: Comments on George Sher." *Philosophia* 36(2): 213–21.

Levy, N. (2012). "A role for consciousness after all." *Journal of Moral Philosophy* 9(2): 255–64.

Levy, N. (2014). "Consciousness, implicit attitudes and moral responsibility." *Noûs* 48(1): 21–40.

Levy, N. and Bayne, T. (2004). "Doing without deliberation: Automatism, automaticity, and moral accountability." *International Review of Psychiatry* 16(3): 209–15.

Madva, A. (in preparation). "Implicit bias, moods, and moral responsibility."

Major, B., Sawyer, P., and Kunstman, J. (2013). "Minority perceptions of whites' motives for responding without prejudice: The perceived internal and external motivation to avoid prejudice scales." *Personality and Social Psychology Bulletin* 39(3): 401–14.

Marino, T. L., Negy, C., Hammons, M. E., McKinney, C., and Asberg, K. (2010). "Perceptions of ambiguously unpleasant interracial interractions: A structural equation modeling approach." *The Journal of Psychology: Interdisciplinary and Applied* 141(6): 637–63.

Mendoza, S. A., Gollwitzer, P. M., and Amodio, D. M. (2010). "Reducing the expression of implicit stereotypes: Reflexive control through implementation intentions." *Personality and Social Psychology Bulletin* 36: 512–23.

Mills, C. (2007). "White ignorance." In Sullivan, S. and Tuana, N. (eds.), *Race and Epistemology of Ignorance*. New York, NY: State University of New York: 11–38.

Moskovitz, G. G. and Li, P. (2011). "Egalitarian goals trigger stereotype inhibition: A proactive form of stereotype control." *Journal of Experimental Social Psychology* 47(1): 103–16.

Nahmias, E. (2006). "Autonomous agency and social psychology." In Marrafa, M., Caro, M., and Ferretti, F. (eds.), *Cartographies of the Mind: Philosophy and Psychology in Intersection*. New York, NY: Springer: 169–86.

Nahmias, E. (2010). "Scientific challenges to free will." In O'Connor, T. and Sandis, C. (eds.), *A Companion to Philosophy of Action*. Oxford: Blackwell: 345–56.

Nelkin, D. (2005). "Freedom, responsibility, and the challenge of situationism." In French, P., Wettstein, H., and Fisher, J. M. (eds.), *Free Will and Moral Responsibility: Midwest Studies in Philosophy*, vol. 29. Ocford: Blackwell: 181–206.

Nier, J. (2005). "How dissociated are implicit and explicit racial attitudes? A bogus pipeline approach." *Group Processes Intergroup Relations* 8(1): 39–52.

Payne, B. K. (2001). "Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon." *Journal of Personality and Social Psychology* 81: 181–92.

Payne, B. K. (2006). "Weapon bias: Split second decisions and unintended stereotyping." *Current Directions in Psychological Science* 15: 287–91.

Payne, B. K. and Gawronski, B. (2010). "A history of implicit cognition: Where is it coming from?" Where is it now? Where is it going?" In Gawronski, B. and Payne, B. K. (eds.), *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications*. New York, NY: Guilford Press: 1–15.

Payne, B. K., Burkley, M., and Stokes, M. B. (2008). "Why do implicit and explicit attitude tests diverge? The role of structural fit." *Journal of Personality and Social Psychology* 94: 16–31.

Pearson, A., Dovidio, J., and Gaertner, S. (2009). "The nature of contemporary prejudice: Insights from aversive racism." *Social and Personality Psychology Compass* 3: 1–25.

Pronin, E. (2007). "Perception and misperception of bias in human judgment." *Trends in Cognitive Sciences* 11: 37–43.

Reznek, L. (1997). *Evil or Ill: Justifying the Insanity Defence*. New York, NY: Routledge.

Samuels, B. (2010). "Unconscious racism at the University of California." <http://www.huffingtonpost.com/bob-samuels/unconscious-racism-at-the_b_491817.html>.

Schlosser, M. E. (2013). "Conscious will, reason-responsiveness, and moral responsibility." *Journal of Ethics* 17(3): 205–32.

Schnabel, K. and Asendorpf, J. (2010). "The self-concept: New insights from implicit measurement procedures." In Gawronski, B. and Payne, K. (eds.), *Handbook of Implicit Social Cognition*. New York: Guilford Press: 408–25.

Schnabel, K., Asendorpf, J., and Greenwald, A. G. (2008). "Understanding and using the Implicit Association Test: V. Measuring semantic aspects of trait self-concepts." *European Journal of Personality* 22: 695–706.

Sellers, R. and Shelton, N. (2003). "The role of racial identity in perceived racial discrimination." *Journal of Personality and Social Psychology* 84(5): 1079–92.

Sher, G. (2006). "Out of control." *Ethics* 116(2): 285–301.

Sher, G. (2008). "Who's in charge here?: Reply to Neil Levy." *Philosophia* 36(2): 223–6.

Sher. G. (2009). *Who Knew? Responsibility Without Awareness*. New York. NY: Oxford University Press.

Sigall, H. and Page, R. A. (1971). "Current stereotypes: A little fading, a little faking." *Journal of Personality and Social Psychology* 16: 252–8.

Stewart, B. D. and Payne, B. K. (2008). "Bringing automatic stereotyping under control: Implementation intentions as efficient means of thought control." *Personality and Social Psychology Bulletin* 34: 1332–45.

Suhler, C. L. and Churchland, P. S. (2009). "Control: Conscious and otherwise." *Trends in Cognitive Sciences* 13(8): 341–7.

Sullivan, S. (2007). "On revealing whiteness: A reply to critics." *The Journal of Speculative Philosophy* 21(3): 231–42.

Rudman, L. and Ashmore, R. D. (2007). "Discrimination and the Implicit Association Test." *Group Processes Intergroup Relations* 10(3): 359–72.

Thalberg, I. (1978). "Hierachical analyses of unfree action." *Canadian Journal of Philosophy* 8(2): 211–26.

Uhlmann, E. and Nosek, B. (2012). "My culture made me do it: Lay theories of responsibility for automatic prejudice." *Social Psychology* 43: 108–13.

Vargas, M. (2005). "The revisionist's guide to responsibility." *Philosophical Studies* 125: 3 399–429.

Vollhardt, J. and Bilewiz, M. (2013). "After the genocide: Psychological perspectives on victim, bystander, and perpetrator groups." *Journal of Social Issues* 69(1): 1–15.

Watson, G. (1975). "Free agency." *Journal of Philosophy* 72(8): 205–20.

Watson, G. (1996). "Two faces of responsibility." *Philosophical Topics* 24(2): 227–48.

Wellman, D. (2007). "Unconscious racism, social cognition theory, and the legal intent doctrine: The neuron fires next time." In Vera, H. and Feagin, J. R. (eds.), *Handbook of the Sociology of Racial and Ethnic Relations*. New York, NY: Springer: 39–65.

Wigley, S. (2007). "Automaticity, consciousness and moral responsibility." *Philosophical Psychology* 20(2): 209–25.

Wolf, S. (1993). "The real self view." In Fischer, J. and Ravizza, M. (eds.), *Perspectives on Moral Responsibility*. Ithaca, NY: Cornell University Press: 151–69.

Zheng, R. (this volume). "Attributablity, accountability, and implicit attitudes."

PART 2

Structural Injustice

2.1

The Too Minimal Political, Moral, and Civic Dimension of Claude Steele's "Stereotype Threat" Paradigm

Lawrence Blum

In 1995, Joshua Aronson and Claude Steele published a paper, "Stereotype threat and the test performance of women and African Americans," that signaled the entry of a powerful psychological insight and ultimately a new research paradigm in the field of educational psychology. Hundreds of subsequent studies have operated within and contributed to this "stereotype threat" paradigm, and in 2010 Steele published a semi-popular book, *Whistling Vivaldi: And Other Clues to how Stereotypes Affect Us*, in which he traced its history and development.

Stereotype threat offers a new way to think about academic and other performance gaps, such as that between men and women in scientific areas, and between whites and blacks in standardized tests and academic subjects more generally. That these racial gaps exist and have persisted over time is a given, although there have also been significant reductions in them in certain historical periods. For example, there was a substantial reduction in the white/black performance gap between the early 1950s and the early 1980s. But the gap has remained fairly steady for the past twenty-five years or so, although the achievement levels of all racially defined groups have improved. Many disciplines have contributed to attempting to understand the reasons for achievement gaps and to finding ways to reduce them. (I will focus on racial groups in this chapter, but stereotype threat has been found to be operative with regard to many different kinds of groups.)

The basic idea of stereotype threat is that as members of groups we are vulnerable to stereotypes of our group that are "in the air," to quote one of Steele and Aronson's most cited articles, not only stereotypes thought to be held by particular persons whom we encounter. Some of these stereotypes directly or indirectly target performance on academic or other tasks, for example, associating our group with lower performance than some other groups, or than a particular comparison group. In such situations, our awareness of the stereotype can result in diminished performance on the task at hand.

Stereotype threat has been isolated as a factor affecting task performance in experimental situations that take the following general form: members of group A, who are stereotyped in the general culture as not doing as well as other groups on a particular task, are given that task. Two experimental scenarios are then contrasted. In one—the stereotype threat condition—the stereotype is made salient to the subject. In the other the stereotype is removed from the situation, so that it is presumed not to be operating in the subject's consciousness. Researchers have consistently found that subjects vulnerable to the particular stereotype do better in the non-stereotype threat condition than in the stereotype threat condition. In the non-stereotype threat condition they do as well as appropriately matched members of the comparison group. If black and white subjects have the same SAT scores, non-stereotype-threatened blacks will perform at the same level as whites on the task at hand, while the stereotype-threatened blacks will underperform the whites.

The stereotype is made salient to the subjects in stereotype threat experiments in several distinct ways. Three important ones are as follows. (1) Subjects are told that their group tends to perform less well on the task to be undertaken than does the control group. (Since most research has been in academic testing situations, I will from here on generally refer to "tests" rather than "tasks.") In this case the subject is provided with the stereotype on the spot, as it were. (2) The subject is told that the test has a certain character that it is plausible to think will cause the subject to worry about a stereotype that her group performs less well on it—for example, a black subject is told that it is a difficult test of verbal or mathematical ability. (3) The subject's racial identity is made salient; for example, by asking the subject to report it on the test, where that identity is often stereotyped as being associated with low performance on this test. In (2) and (3), the manipulation is plausibly taken to make salient to the test-taker the stereotype that blacks have less ability in verbal and mathematics areas, even though the stereotype is not explicitly mentioned.

The non-stereotype-threat or "stereotype-neutral" condition is created by various manipulations. In one, from Steele's early experiments, the black subject is told that the test is not being used to test for ability but is being used to test the test itself. The presumption is that if the subject does not think her individual score is at issue, she will not labor under a threat of being stereotyped. A second way of neutralizing the stereotype is to tell the subject that members of her group perform as well on this particular test as do members of other groups, and specifically as well as the control group.

As to the mechanism by which the presence of stereotype threat depresses performance, researchers have not reached consensus. Steele says that it operates through the subject's concern not to confirm the stereotype, or worry that she will confirm it. But how exactly does that state of mind cause diminished task performance? Maass and Cadinu, in a 2005 overview of the state of stereotype threat research, summarize some of the theories in play: (1) concern about the stereotype causes anxiety, which interferes with focus on the test itself; (2) concern causes intrusive thoughts related to the stereotype, which interfere with test focus; (3) the subjects are more cautious than they would otherwise be in fear of confirming the stereotype, so they answer fewer questions, causing underperformance and so ironically defeating the very point of the caution; and (4) the subjects lower their self-expectations in a not-entirely-conscious response to the stereotype that they will not do as well; they in some way internalize the stereotype that they will not perform as well or do not have as strong abilities as the control group. The reduced self-expectation then reduces the effort put into the task.

Steele himself does not have a settled view on the mechanism issue, but he tends to favor interference with cognitive processes: that is, options (1) and (2). He repeatedly emphasizes that stereotype threat does not operate by the subject's internalizing the stereotype, causing self-doubts or a sense of incapability: that is, option 4. On his view, the subject does not suffer from a diminished confidence that she can perform the task, nor does she lower her expectations of herself in doing it. Rather, on Steele's view, she tries as hard as she can, and as hard as the other test-takers of other identity groups; but her concern not to confirm the stereotype by doing less well than others paradoxically causes her to perform less well.

To some degree, Steele's position denying the internalization of stereotypes as a mechanism of underperformance in stereotype threat is merely terminological. He does not flatly deny that underperformance can be caused by internal self-doubts, but he wants to reserve the term "stereotype threat" for a process distinct from internal self-doubt. Other researchers, such as Maass and Cadinu, and the sociologist Douglas Massey (whom I will discuss in Section 9), count any underperformance caused in some way by the presence of a heightened stereotype as a manifestation of stereotype threat.

However, one important feature of Steele's view inclines him sometimes to go beyond the terminological move, to come close to denying that internal self-doubt is operating at all in task performance situations. This feature is that his subjects are very successful students and strongly identify with the domain of the task at hand. Steele's initial research was entirely on students at the very top selective colleges, such as the University of Michigan and Stanford. Steele claims, with some plausibility, that these accomplished black students are those least likely to have self-doubts about their own capability. Their academic success shows them that they are indeed capable and protects them against self-doubt, while at the same time rendering them vulnerable to stereotype threat as he defines it. Because of this focus, one could say that stereotype threat as Steele sometimes understands it is a phenomenon only of those without self-doubt in the domain in question.

By contrast, students who are not concerned with how well they do in a task domain, perhaps because they have not been successful in that domain previously and so have disengaged from it, are relatively immune to stereotype threat. When the stereotype of their group as academically or intellectually incapable is made salient to them, this has little effect on how well they do on the tasks at hand, because they are not invested enough in the task to care about the stereotyping of their group in that domain. (Note that this disinvestment in the task, or the domain of the task, can itself be partly caused by *prior* stereotype threat. Such disidentification is a long-term effect that Steele indeed notes; but his work is focused on the short-term effects that are more easily measured.)

What do stereotype threat researchers propose as corrective or preventative interventions that mitigate or eliminate stereotype threat in educational environments? Steele mentions several. In one, white professors established trusting relationships with black and Latino students—for example, establishing high academic standards and credibly expressing confidence that the black or Latino student can meet those standards.[1] A second involves dorm-based discussions among students of various racial groups about challenges faced by all of them; the black and Latino students see that their own struggles are shared by other students, and feel more comfortable at the university as a result (Steele, 2010: 167). More generally—a point made by Maass and Cadinu—if the student's racial identity is made less salient, for example by some other identity ("students at

[1] Steele cites the findings of Massey et al. that stereotype threat is almost entirely absent among black and Latino students when the professor is black, though he says that that outcome could (also) be an effect of a critical mass of black and Latino students in these instructors' classes (Steele, 2010: 159).

college X," "struggling college students") coming to the fore, the student will be less vulnerable to stereotype threat (Maass and Cadinu, 2003: 250ff). Another general approach draws on the US Supreme Court's majority opinion in the 2003 affirmative action case—that the presence of a "critical mass" of a vulnerable group's members helps members of that group feel a sense of belonging at the college, especially if coupled with messages from teachers and college officials that the group in question is a welcome and valued presence.[2]

In this chapter I do not question the essential soundness of the stereotype threat paradigm, nor its essentially progressive thrust as a way to enhance the performance of vulnerable racial groups and shrink academic gaps. Stereotype threat research, and research on implicit bias as well, have done a tremendous service to public thinking about race by making it clear that the reigning color-blind ideology is wildly out of line with the facts on the ground. Race very much still matters, and is pervasive in the barriers facing racial minorities and in the attitudes and behavior of actors in a position to affect the life situations of racial minorities. From a political point of view the stereotype threat paradigm possesses the remarkably attractive feature of both highlighting racial barriers faced by racial minorities yet not highlighting the carriers of the stereotypes as particular white people; the emphasis is on stereotypes being "in the air" rather than held by particular white people whom the racial minorities encounter. Implicit bias is significantly different in this regard since it conveys the message that well-meaning white people do carry prejudices and stereotypes they are not aware of and would disavow were they to come to recognize them.

I will criticize the stereotype threat paradigm and research carried on within it on two general grounds. First, while it helps deal with one piece of the undoubtedly important issue of racial disparities in achievement, it is quite narrow when viewed in the broader context of education, including moral and civic education, concerning stereotypes and stereotyping more generally. Stereotype threat research does not look into ways that students in stereotype-vulnerable groups can be equipped with the intellectual tools to recognize, examine, and criticize stereotypes in general and stereotypes of their group in particular—an endeavor with moral and civic significance. For example, it does not point the way to

[2] Grutter v. Bollinger et al., 539 U.S. 306 (2003), O'Connor opinion. Steele characterizes the range of approaches to reducing stereotype threat suggested by research in this way: "[E]stablishing trust through demanding but supportive relationships, fostering hopeful narratives about belonging in the setting, arranging informal cross-group conversations to reveal that one's identity is not the sole cause of one's negative experiences in the setting, representing critical abilities as learnable, and using child-centered teaching techniques." (181)

helping students of all groups to be able to distinguish stereotypes from valid generalizations. Finally, the stereotype threat paradigm does not encourage or draw on the concern members of such groups do or could feel for one another to enlist them in a collective project of challenging stereotypes that harm their group. These are all prime candidates for moral and civic education regarding stereotypes.

My second criticism is that the stereotype threat paradigm gives insufficient credence or at least attention to systemic injustices against stereotype-vulnerable groups. One reason for teaching the difference between a stereotype and a valid generalization is that valid generalizations are essential to recognizing and characterizing systemic racial injustices; they should not be avoided for fear of the taint that appropriately attaches to "stereotyping." I also argue that Steele's keeping at arm's length the idea that members of stereotype-vulnerable groups are prone to internalize negative stereotypes of their group contributes to his failure to include as a core part of his analysis the systemic and multifaceted nature of race-based injustice. Steele thus shares with other psychology-based approaches to racism and racial inequality (such as implicit bias) a tendency to sideline more structural and systemic causes of inequality, including educational in equality.

1 What are Stereotypes?: Distinguishing Valid from Invalid Generalizations and Associations

Let me begin with a lack of clarity in what "stereotype" is taken to mean in the stereotype threat research paradigm. It seems to involve associating a social group with a particular attribute related to performance on specific types of tasks. The attribute is generally a negative one; that is, one that links the group in question to a less-than-equal or less-than-adequate capacity for performing the task in question.[3]

I say that the notion of stereotype involves an "association" of a group with an attribute, rather than a *belief* that the group possesses the attribute. The type of link that stereotype threat research assumes between the group and the attribute

[3] But the attribute itself does not have to be negative. Some stereotype threat researchers have looked at the way that a stereotype of Asians as good at mathematics enhances their performance compared to other groups and compared to Asians in a non-stereotype threat condition. This enhanced performance obviously cannot be accounted for by Steele's favored mechanisms, which interfere with maximal cognitive functioning.

does not seem to require the cognitive robustness of a belief; "association" seems more appropriate.[4]

Stereotype threat researchers do not generally note whether the associations in question are warranted or rationally appropriate. It may be that their use of "stereotype" to mean such an association whether warranted or not has become standard in psychology; there is some evidence for this.[5] But the ordinary language usage of the word "stereotype" is evaluatively valenced, and does, rightly I think, imply the importance of the distinction between warranted and unwarranted associations or generalizations. "Stereotype" normally implies something wrong with or about the association being made between the group and the attribute.[6] What constitutes that wrong is perhaps a matter of dispute, but I would suggest at least two elements: (1) that the stereotype involves or implies a false generalization about a group, sometimes in the form of an overgeneralization; (2) that stereotypes as mental representations are more resistant to counterevidence than are mere false generalizations. In being evidence-resistant, stereotypes also involve other epistemic wrongs or bads. For example, they skew perception of members of the target group so that counterstereotypic evidence—members of the stereotyped group not behaving in accord with the stereotype—is not noticed in the first place; or it is noticed but does not dislodge the stereotype of the group in general (e.g. because the counterinstance is seen as an "exception" or as not typical of the group). Whether the ordinary use of the word "stereotype" actually has these implications is less important than whether we can distinguish generalizations or associations that have these features from those that are valid, true, sound, and the like, and whether this distinction is worthy of note.

To the retort that Steele would not care about the distinction between valid and invalid associations because it is the mere association itself, valid or not, that causes stereotype threat, I would reply that I would like to see what would happen in the experimental situation if the experimenter highlighted that

[4] Gendler's notion of "alief" might also be an appropriate way to characterize the cognizer's relation to a stereotype that she holds (Gendler, 2008).

[5] Here are examples of warrant-neutral definitions of "stereotype" in social psychology: "Stereotyping is the process of ascribing characteristics to people on the basis of their group membership" (Oakes, Haslam, and Turner, 1994). "Stereotypes are beliefs about the characteristics of groups of individuals" (Stangor, 2000). (Cited in Blum, 2004: 256.)

[6] It is significant that when Steele talks about "disproving a stereotype," he does not generally mean demonstrating that it is a false or overstated generalization about a group, but rather that it is not true of the individual agent in question. See 111–12, for example, "Disproving a stereotype is a Sisyphean task; something you have to do over and over again as long as you are in the domain in where the stereotype applies."

distinction in her instructions to the high achieving subjects in something like this way: “On average, blacks do not do as well on this sort of test as whites do. However, what is true of averages is not true of all subgroups or individuals. Your academic records are far above the average for black students, so what applies to black students on average regarding achievement likely does not apply to you. We are confident that you will do as well on the test as the white students.” These instructions report the association with the subject’s racial group as a whole, but show the subjects why that association is not likely to hold in their case. Would the high-achieving black students still experience stereotype threat in such a scenario? Indeed, might not such instructions be experienced as creating something like the stereotype-neutral condition (discussed previously, 148f). If so, this suggests that true generalizations about the average underperformance of their group can be made salient to test-takers in a way that does not necessarily engage mechanisms that produce stereotype threat.

2 Imprecision in Content of Stereotypes

Steele and Aronson are also often imprecise regarding the content of stereotypes. In one of their important articles, for example, they refer to “negative stereotypes about [a target group’s] mental ability” without explaining what is meant (Steele and Aronson, 1998: 401). On other occasions, they speak of what they imply to be the *same* stereotype in terms of “low performance.” But these two formulations—“(negative stereotype of) mental ability” and “low performance”—are not equivalent. Low *performance* can be caused by all sorts of things other than low *ability*.

This conflation is likely to have a bearing on stereotype threat. To be stereotyped as having low performance is less demeaning than to be stereotyped as having low ability, since some of what causes low performance may be out of one’s control and so not reflect the fundamental deficiency that “low ability” does. It is at least plausible to think the two stereotypes would or at least could have different effects; and so one would want to clearly distinguish between them.[7]

This is to say that when Steele talks vaguely about the stereotype that links blacks to performance or ability, several distinct and importantly different things could be meant by this.

[7] An important further distinction within “ability” is between innate versus developed (conceptions of) ability. These could produce differing stereotypes that might have differential effects on stereotype threat.

3 The Epistemic, Moral, and Civic Harms of Stereotypes and Stereotyping

Steele's expressed educational interest in stereotyping lies in its role in causing stereotype threat. I have mentioned two educational interventions regarding stereotypes—learning the difference between evidence-resistant overgeneralizations (stereotypes) and sound and factual generalizations, and learning to be aware of important ambiguities and unclarities in the stated content of stereotypes—that can plausibly be thought to have some effect on vulnerability to stereotype threat.

But the educational import of understanding these things about stereotypes is much broader, touching on moral, civic, and intellectual aims at the heart of the educational enterprise. We may start by recognizing the range of harms that being stereotyped causes to stereotyped groups and their members that valid generalizations do not, and that similar but different stereotypes cause different harms or degrees of harm.

Without attempting to be comprehensive, let me mention some of these harms. Stereotypes produce false views of groups and of individual members of those groups. Depending on the content of the stereotype (such as the ones above), those false views can be demeaning and disrespectful. Some can be more demeaning than others. A stereotype that your group is inferior in mental capacity to another group (either race or gender based, for example) is more demeaning than one that says that your group underperforms another group, leaving the explanation of this unspecified. (This point was made above, but only in relation to different impacts on stereotype threat.)

Stereotyping involves failing to see or to appreciate the internal variety and diversity within a group. Properly understood, true group generalizations do not involve this epistemic fault. A student who understands that group A's average performance on a test (or the percentage of the group that reaches some specified and informative standard) is greater than group B's (or the percentage of group B that meets that standard) is able to recognize, and should be taught, that there is likely to be significant variation in performance within both groups A and B. Stereotyping homogenizes groups, tending to mask their internal variety (though not necessarily denying exceptions that leave the association in place).

This intellectual distortion involved in stereotypes has moral and civic implications. Failure to see internal diversity in a group is a failure of recognition of the character of the group and of its members. For example, if a student holds the stereotype of Asians and Asian Americans as being hard working and superior students, she will often fail to recognize Asian Americans who are not. They will not register with her or, if they do, she may even stigmatize them for not living up to her stereotype of that group. More generally, stereotypes are serious obstacles to seeing members of other groups as individuals, and thus to according them the respect due each person as an individual.

In masking the internal diversity, stereotypes overstate the difference between that group and other comparable groups. If the student holds a stereotype of the form "As aren't serious students" and "Bs are serious students," she will experience As and Bs, both as groups and as individual members of those groups, as much more different from each other than they really are. Stereotypes also often involve a moral distancing from the stereotyped group and its members—seeing them as very much unlike and separate from oneself—and some do so more than others.[8]

All these are ways that stereotyping blocks the appropriate regard and respect which students should learn to have for members of groups other than their own. They apply to all stereotypes—those that attribute positive characteristics to the group, as well as those that attribute negative ones. But negative stereotypes involve further forms of misrelationship with others. Holding a stereotype of a group as stupid, lazy, or alcoholic is demeaning or insulting to the group and its members, thus violating the norm of respect for others that should be part of any moral education program. An accurate, well-founded generalization that group A has higher rates of treatment for alcoholism, or that group C's average performance on a standardized test is lower than group D's average, do not carry this whole group disrespect if the cognizer recognizes it as a valid generalization and does not confuse it with a stereotype.

These forms of misrelationship and misunderstanding of others involved in stereotyping have civic as well as moral and epistemic significance. Insofar as many stereotypes have as their object important social groupings in society—race, ethnicity, sexual orientation, nationality, gender, and the like—stereotyping presents a serious obstacle to developing the civic capacity of engaging respectfully with those who are different from oneself in the service of promoting a common good.

I have illustrated only some of the ways that stereotyping involves cognitive and moral misrelationships with others.[9] My point is to show the educational importance of helping students—all students, not only those in stereotype-vulnerable groups—learn how to recognize stereotypes, how to analyze their

[8] The previous five paragraphs are a brief summary of Blum (2004). See also Anderson (2010: ch. 3).

[9] For more on the morality of stereotypes, see Blum (2004); Stephan (1999).

content, how to differentiate them from valid generalizations. Acquiring the intellectual tools to do so has important civic, moral, and personal benefits.

4 Failure to Promote and Draw on Group Consciousness among Stereotype Threatened Groups in Stereotype Threat Research

One further moral benefit of learning about stereotypes bears mentioning because it is strikingly absent in Steele's discussion. That is, that it helps a member of the target group to regard her entire group and its members in a more positive light—as being just as capable as members of other groups (Blum, 2012: 110–12). Steele and his associates focus only on the benefit to each individual in a stereotype-vulnerable group of being able to work up to the level of her developed capabilities. Their paradigm does not encourage members of vulnerable groups to be concerned *about one another* in relation to their shared vulnerability to the stereotype. It thus also fails both to draw on and to support a sense of shared identity in the face of this educational challenge facing fellow members of the group. The more comprehensive education regarding stereotypes that I am recommending involves both broader educational growth for the individual and also a morally significant greater concern for and appreciation of the capabilities of members of her group.

I am not arguing that if presented with a program of broader education about stereotypes Steele would reject it as unimportant, claiming that the only significance to stereotyping is its producing stereotype threat. I argue only that he shows almost no actual recognition of these wider educational aims in what I have read of his work and especially his book *Whistling Vivaldi*. He does not situate his discussion of stereotypes in a broader context of other educational concerns about them. The only harm ever mentioned with respect to stereotypes is their contributing to purely academic performance gaps. In addition, the basic intervention strategy is for third parties to neutralize the stereotype so that they do not impinge on the subjects in the target group. The interventions do not aim to help the subject learn to defang the stereotype herself by becoming convinced that the stereotype is wrong, for example, through deepening her understanding of stereotypes and valid generalizations.

Because of the tremendous influence of the stereotype threat research paradigm, I also worry that its silence on these broader educational issues will contribute to funders of educational research failing to recognize their importance.

But my critique goes a bit beyond this silence. At one point Steele seems to offer a reason for not trying to convince stereotype-vulnerable persons that a familiar stereotype of their group is false. With respect to the stereotype that women are not good at mathematics, he says, "[E]ven if we could convince them [of the falsity of the stereotype], it was doubtful we could convince them that other people didn't believe the stereotype, broadly held as it is. And if we couldn't convince them of that, they could still worry that their test performance would cause other people—the experimenter perhaps—to see them stereotypically" (Steele, 2010: 38).

This seems too quick a response. Perhaps if someone were truly firmly convinced of the falsity of a given stereotype of her group, she would be less concerned about those not of her group holding that stereotype. As far as I can see, this hypothesis is not tested. But it is not implausible to think that people who themselves differ in the degree to which they are convinced of the falsehood of a stereotype of their group might differ in their degree of vulnerability to what they perceive as the investment of others in that stereotype.[10]

Implicit bias seems to me vulnerable to a criticism similar to the loose use of "stereotype." Implicit bias refers to as "bias" an association between a group and a negative or positive characteristic. But whether that association is well-founded should affect whether we call it biased or not. If someone associates black students with academic underperformance because she is aware that by many significant measures of performance, blacks perform less well than whites, this does not seem to me a "bias." It might plausibly be called a "bias" if the association is attributed to the group as an implied universal generalization, but not if its actual degree of factual basis is appreciated. Moreover, even if accurate, it would be wrong to base expectations about a particular individual student on it. But it is one thing to criticize the mere association itself (e.g. by calling it a bias) and another to criticize certain further cognitions or expectations made on the basis of it. This distinction is noted by most implicit bias researchers. However, they still tend to see the association itself as a bad thing. I am suggesting that the association itself is not a bad thing when it reflects a well-founded generalization, only when it does not.

5 Role of Accurate Generalizations in Diagnosing and Correcting for Unjust Disparities

A different but also morally, civically, and epistemically important reason for distinguishing true generalizations from stereotypes, and for helping students to

[10] See further discussion of the issue of being convinced of the falsehood of a stereotype of one's group (Section 10, 169f).

recognize and appreciate the importance of the distinction, is that stereotypes often play a role in rationalizing group inequalities. True group generalizations, by contrast, help us diagnose the existence and causes of such disparities so that we can try to change them if they are unjust—and engaging with injustice is a key component of civic education.

If all associations between a group and an attribute are regarded as suspect because they may contribute to stereotype threat, we miss the role that true generalizations play in understanding systemic and structural injustice.[11] That blacks and Latinos on average underperform whites and Asians by substantial margins on a range of educational measures is, or points to, a serious systemic injustice in our educational system. This injustice is expressed in various different true generalizations—for example, that blacks and Latinos drop out of high school at greater rates than whites and Asians; that among high school graduates a smaller percentage of Blacks and Latinos attend four-year colleges than do whites and Asians; that among college matriculants, a higher percentage of Blacks and Latinos fail to complete their four-year degrees than do whites and Asians (Farkas, 2008).

Unless these disparities are kept firmly in mind, we and our students will not be prompted to inquire into their sources, which are multiple, and which generally implicate structural features of society—degree of poverty, especially concentrated neighborhood and school poverty; increasing economic inequality that both intensifies inequality in access to quality education and also renders the powerful and politically well connected less aware of the plight of the disadvantaged; inadequate resources provided to schools with large populations of black and Latino students; segregation of schools by race and class; and many more such considerations (Ladd, 2011; Anderson, 2010; Duncan and Murnane, 2011). I am not attempting to provide an analysis of the myriad causes of educational disparities, but simply making the elementary point that such an analysis must begin with a cataloguing of those disparities themselves. The ability to recognize and assess group generalizations is a key civic capability that students should learn, hopefully beginning in high school or even before.

This is a reason to resist using the word "stereotype" for all associations between blacks and low academic performance, because the danger that the stigma often associated with that term will make it difficult to talk frankly about existing racial disparities. Several education researchers have noted ways

[11] Stereotypes as they appear in stereotype threat can be regarded as in a sense systemic in that they are widely shared and often (not always) viewed as "in the air." But they are not regarded as structural, or part of a structure that creates disparities. (See discussion in Section 6.)

that school staffs' reluctance to take note of race-based patterns of behavior (e.g. rates of school suspension) and academic performance for fear of stereotyping members of certain student groups disables them from being able to inquire into causes of and solutions for problems that affect those student groups (Pollock, 2005; Schofield, 1989). This is a further reason to build into our educational response to stereotype threat an appreciation of the distinction between a stereotype and a sound generalization, and of the role of true group generalizations in analyzing the causes of these disparities. Steele is aware of these disparities and occasionally refers to them in his writing (e.g. 2010: 158; Gates and Steele, 2009: 257). But his loose use of "stereotype" is an obstacle to a frank recognition of them and their importance in diagnosing injustice, and they play a peripheral role in his concern with group disparities.

6 How Stereotypes Support and are Supported by Unjust Social Structures and Systems

The stereotype threat perspective sidelines structural injustice in another way also. Regarding actual stereotypes—not accurate generalizations that may be conflated or confused with them—the stereotype threat perspective downplays the ways stereotypes are themselves embedded in structures of group inequality. Recent discussions of racial inequality, such as Elizabeth Anderson's *The Imperative of Integration*, Glenn Loury's *Anatomy of Racial Inequality*, and L. Bobo and M. Massagli's "Stereotyping and urban inequality," have shown how inequality tends to generate and reinforce representations of disadvantaged groups as inferior to advantaged ones (Anderson, 2010; Loury, 2002; Bobo and Massagli, 2001; see also Shelby, 2003). For example, although black students' underperformance in American schools can be traced to a complex set of social/environmental factors, such as mentioned previously, the fact of underperformance alone often leads members of all groups to think of black students as having lesser abilities—to stereotype them in that way.

Also, conversely, in often subtle ways, representations/stereotypes of disadvantaged groups as inferior helps to sustain structures in which those groups are kept in a disadvantaged position. Elizabeth Anderson nicely lays out several mechanisms by which this takes place, that lie outside standard understandings of racial animus and prejudice but do include stereotypes. She employs the notion of "stigma"—a socially salient dishonored identity—to argue that both inequality and segregation reinforce stigma toward blacks, often in interaction with cognitive distortions of the sort consistent with the findings of implicit bias research (Anderson, 2010: ch. 3). For example, "system justification bias" inclines people to interpret their social world as just, and thus to attribute favorable characteristics to the advantaged and unfavorable ones to the disadvantaged (Anderson, 2010: 46; Jost, Banaji, and Nosek, 2004). Thus part of our concern with stereotypes should be with how they support as well as are generated by structures of injustice.

But Steele presents stereotyping primarily as a free-floating cognitive distortion, untethered from these structures of inequality. His focus on high achieving students is in line with this approach. As a group, black students at elite colleges have in a sense largely (though not entirely) escaped the main systematic structures of racial disadvantage. While black students as a whole are, on the average, subject to lower teacher expectations, the effects of concentrated poverty, lesser familial resources, weaker teachers, and the like, Steele's particular group, some of whom may have been subject to some of these disadvantages at earlier stages, have made it to a place where they are no longer kept from success by those factors. Their somewhat lesser performance at the college level compared to whites must be placed in the context of their exceptional successes in achieving an academic record at prior levels that propels them to the pinnacles of higher education institutions. The stereotype threat-blocking, and apparently remarkably effective, interventions that Steele suggests for these students generally involve some form of getting such students in touch with their achievements and the abilities and traits of character that have led to them, as a way of keeping the stereotypes at bay. The interventions thus implicitly remind the students that the disadvantaged situation of other blacks does not apply to them, or to nothing like the same degree. Other interventions (such as convincing the students that a given test is not ability-diagnostic) take the stereotype out of the picture, allowing the students' superior developed abilities to take their natural course. Steele's approach relegates structural injustice to the background because he focuses on subgroups within the disadvantaged group who have already largely escaped or transcended the factors disadvantaging the groups as a whole.

But the logic behind these interventions renders them significantly less applicable to a much greater number of other blacks for whom the myriad factors harming their achievement remain in effect. Thus if we think of the stereotype threat paradigm as taking on one significant source of educational inequality, it is important to recognize how vast is the domain of factors *not* thereby encompassed. I have mentioned some features of that domain previously, but let me focus more closely on two of them for purposes of illustration. One is the widespread phenomenon of so-called "ability grouping" as a principle of organizing the assignment of students to classes and selecting curriculum and pedagogy for students so assigned. High schools with racially mixed populations generally have advanced placement and honors classes, and lower level classes. Karolyn Tyson remarks that "the image of overwhelmingly black lower-level and overwhelmingly white advanced classes . . . sends powerful messages to students about ability, race, status, and achievement" (Tyson, 2011: 8).[12] This form of organization of learning negatively affects many black students (Tyson, 2011; Oakes, 2005). It is a type of obstacle different in character from stereotype threat among the intensely academically-identified high achievers, affecting the much greater number of students not placed in the most advanced classes.

A second type of factor is school segregation by race and class—a common phenomenon in American school systems. Schools with a high concentration of poor and black or Latino students face a daunting range of challenges to delivering high-quality education—working conditions that tend to drive away the most qualified teachers (with the exception of an exceptionally dedicated minority); unattended to health problems; economic insecurity in families and consequent stresses of home life; frequent moving in search of better economic opportunities; the multiplier effect of high concentrations of these problems within the individual school; and others (Blum, 2015; Rothstein, 2013). These deficits are only weakly related to stereotypes. Yet under more favorable circumstances, many students in these schools could flourish educationally.[13]

[12] Tyson is showing how stereotypes are generated by the structures that themselves contribute to unequal educational outcomes. The stereotypes may then be a further contributory factor to those disparities. But Tyson's point is that they themselves are generated by the structures. And she is also saying, contrary to Steele, that they operate by lowering students' race-based self-expectations (Tyson, 2011).

[13] Steele sometimes remarks on these and other factors besides stereotype threat that contribute to education disparities. For example, he refers to black students' skill deficiencies, clearly differentiating this factor from stereotype vulnerability (Steele, 2010). My argument does not depend on claiming that he denies these factors, only that he fails to accord them adequate importance, and to connect them to moral and civic education relating to stereotyping. But it is worth noting that Steele sometimes slips into ways of talking that come very close to implying that stereotype threat is the sole cause of educational disparity. For example: "I don't want to imply that the pressure of identity threat and its cumulative impact on African Americans, even in intellectual areas, is so unmitigated and foreclosing as to allow only a few individual successes in these areas. There are clearly many such successes and many factors that can mitigate this threat for individuals" (Steele, 2010: 133). Clearly, the many successful African Americans have avoided being decisively held back not only by stereotype threat but by the many other barriers to African American success, some of which I have mentioned, but which disappear in this passage.

7 Problems with Steele's "Everyone is Subject to Stereotype Threat" Strategy

Two further features of Steele's analysis divert attention from the connection I am highlighting between stereotypes and structures of social injustice. One is his claim that every group—not only disadvantaged groups—suffer from or can suffer from stereotype threat.[14] Some stereotype threat paradigm experiments involve white students' underperformance on a mathematics test about which they are told that Asians tend to do better than whites, compared to a control group of whites who are not told this. Steele infers from this that stereotype threat is not confined to disadvantaged groups but holds for any group that can be made vulnerable to a stereotype that casts their abilities as inferior to some other group.[15]

Yet even if that is so, disadvantaged and subordinate groups are much more broadly vulnerable to stereotype threat in real life than are dominant groups.[16] The stereotype of whites as not being good in mathematics is not "in the air" in the way that anti-black stereotypes are. It can be found only in unusual real life situations, or produced artificially in experimental situations (although to be sure it draws on a cultural stereotype of Asians as good in mathematics), so blacks will face negative stereotypes much more frequently and in more kinds of life situations (e.g. interviewing for a job) than will whites face the mathematics-related stereotype.

So underperformance has a completely different significance for advantaged and disadvantaged groups. In one it is interwoven with a social system of inequality, in the other it is not; the former is a much more serious justice concern. While stereotype threat analysis does not explicitly deny this point, it

[14] This point differs from one that some stereotype threat researchers make, that disadvantaged groups other than race- and gender-based ones—older people, for example—suffer from stereotype threat (Maass and Cadinu, 2003: 270). Steele et al.'s conception of disadvantaged groups is a broad one. The point in the text applies to non-disadvantaged groups.

[15] "Terms like 'yuppie', 'feminist,' 'liberal,' and 'white male' can all call up unflattering images" (Steele and Aronson, 1998: 401). Another interesting example is male students performing worse on a test of "social intelligence" than female students; the exact opposite result on the very same test is produced by labeling the test "logico-mathematical intelligence" (Maass and Cadinu, 2003: 269).

[16] Even if dominant groups are vulnerable to stereotype threat, and even if the degree of underperformance is the same, it is possible that the mechanisms driving the underperformance in the case of disadvantaged and subordinate groups, compared to dominant or advantaged groups, are not the same. Maass and Cadinu suggest that the underperformance in the subordinate group may stem from lowered self-expectations caused by stereotype salience, while in the dominant group case it might be a different kind of interference with cognitive processes, more along the lines of Steele's view (Maass and Cadinu, 2003: 251). As mentioned previously, Steele tends to reject the former version of the mechanism of underperformance.

tends to mask it. Steele may indeed see this masking as part of a political strategy to garner attention to stereotype threat. If everyone can experience stereotype threat, then everyone can relate to it, and a concern with it in the case of subordinate groups can thus be framed as part of a universal concern rather than a particularized concern on behalf of those groups. In the US in recent decades, and very much also since Obama's election in 2008, it has been difficult to muster popular or legislative support for programs targeted toward disadvantaged racial groups. The eminent sociologist William Julius Wilson is famous for having called on those concerned with racial justice to de-emphasize specifically racial concerns and to look for universal race-neutral programs that would disproportionately benefit blacks and Latinos (Wilson, 2001).

This "universalizing" strategy is expressed in the book flap description of Steele's *Whistling Vivaldi*: "But because these threats, though little recognized, are near-daily and life-shaping for all of us, the shared experience of them can help bring Americans closer together." Without here assessing the strategic issue, it is important to note that there is a cost to sidelining the moral differences between stereotyping of subordinate and of dominant groups. Attention to racial injustice is an urgent moral and civic matter that universalizing approaches can go only so far in addressing, and that Steele's approach may risk masking. As the philosopher Tommie Shelby remarks in discussing Obama's embrace of universalizing programs and avoidance of initiatives to address racial injustice directly: "This strategy would intentionally obscure the morally important difference between creating more opportunity for all and remedying the effects of past racial injustices" (Shelby, 2011: 104).

I note that implicit bias research does not go in for this misleading symmetrizing. For implicit bias the targets of prejudice are vulnerable, subordinated, or devalued groups—racial minorities, the aged, women, gay people—not advantaged groups.

8 Internalized Stereotyping and Systemic Injustice

A second feature of Steele's view also diverts attention from the connections between stereotype threat—or the underperformance that stereotype threat is meant to explain—and systemic and structural racial injustice. That is Steele's rejection, mentioned previously, of the idea that some blacks may have come unwittingly to cognitively invest, at least partially, in a stereotype about themselves as intellectually less capable than other groups, especially whites; that is, that they have (partly) internalized the stereotype. As I mentioned earlier, Steele is not entirely consistent in what he says about this. At his most careful, he does not deny that this process is operating, but wants only to affirm that the phenomenon he wants to call "stereotype threat" is distinct from it (Steele, 2010: 93). However, Steele's not infrequent and often dismissive references to internalized racial self-doubts as "psychic damage" or internalized self-hatred show that he is far from embracing the phenomenon itself as part of the explanation of black underperformance.[17] This distancing is reflected in an important change in terminology in Steele's research. In their first experiments Steele and Aronson employed "stigma" rather than "threat." As mentioned in Section 6, "stigma" implies a socially embedded dishonored identity—something that is not confined to a specific situation, but relates to the general standing of the group in society. By contrast, "threat" allows for the situation-specificity that, as we have seen, Steele heads toward, even if not entirely consistently. But as Anderson and Loury argue, stigma is a better description for the ways that stereotypes operate in relation to subordinate groups (Anderson, 2010; Loury, 2002). A stigmatized group is not merely threatened with the worry that people will look down on them in specific stereotype-triggering contexts. The socially dishonored identity pervades the lives of stigmatized groups. In light of this it would be very surprising if some substantial percentage of the stigmatized group itself did not to some extent internalize the stereotype of the group as deficient in some important respect, resulting in a degree of self-doubt.[18]

Both Anderson and Loury show how stigma contributes to pervasive inequality. Diverse social actors drawing, consciously or not, on stigmatizing representations of blacks may leave neighborhoods they feel are becoming "too black," may show blacks looking for housing a narrower range of options, may expect less of black students, may prescribe inferior medical treatment to black patients,

[17] Some examples of Steele's dismissive or at least distanced references to stereotype internalization: "The psyche of individual blacks gets damaged, the idea goes, by bad images of the group projected in society—images of blacks as aggressive, as less intelligent, and so on. Repeated exposure to those images causes these images to be 'internalized,' implicitly accepted as true of the group and, tragically, also perhaps of one's self" (Steele, 2010: 46). (The tone of this quote, and more importantly, the subsequent argument, show that Steele treats this view as a competing explanation of underperformance to his own, and one that his research shows to be false.) Speaking of himself as a new graduate student vulnerable to stereotype threat, he says "I don't believe I had this view [not being certain he belonged at the institution where he was studying] because of some 'psychic damage' that perhaps grew out of my experience of race in the United States, for example" (Steele, 2010: 163). And in an interview with Henry Louis Gates, Jr: "I have spent a lot of time trying to make something clear. I'm not talking about internalized self-doubt and self-hatreds—the classic 'psychic damage' in Daryl Michael Scott's phrase" (Gates and Steele, 2009: 252).

[18] I am not criticizing Steele for the terminological shift from "stigma" to "threat" as it more accurately reflects the target of his research program as it developed. I am only noting that something is lost in this shift with respect both to the terminology and the research program.

may prefer white job candidates to equally qualified blacks, and so on.[19] There need not be racial animus behind this discriminatory treatment—only shared demeaning representations of the group in question. Indeed, it is perfectly in line with Steele's understanding of stereotype threat to recognize that stereotyping can operate outside of actual racial antipathy and prejudice. But because stigma is pervasive in society it is likely that some members of the stigmatized group will internalize it, at least to some extent. It is this point that Steele's view denies or relegates to the sidelines.

9 The Implausibility of Steele's View on Internalized Stereotyping

That Steele's view on internalized racial self-doubt is counterintuitive is suggested by an important study, by Douglas Massey and his co-authors, of the social origins of students at selective colleges. Massey et al. utilize two criteria of "vulnerability to stereotype threat," in an attempt to discern the causes of black and Latino underperformance at these colleges. One is an exceptional self-consciousness about what their teachers think of them; the other, that they do not regard themselves as very good students. That is, Massey et al. count students having doubts about their abilities as "vulnerability to stereotype threat," and they investigate the extent to which such self-doubts affect these students' performance (Massey et al., 2003). This seems a very reasonable approach; but, although the authors imply that they are working within Steele's research program, theirs is the very scenario—i.e. internal self-doubt—that Steele rejects as being part of stereotype threat.

Polling data confirms the idea that blacks to a not insignificant degree hold some of the same devaluing associations or images of themselves that whites do of them. Massey et al. report a study of the racial attitudes of their college student subjects that finds that blacks score blacks as the "laziest" of any of the four racial groups (both higher than they scored other groups and higher than other groups scored them on laziness, and also less intelligent than they ranked the other three racial groups: Asian, white, Latino). "Thus, the stereotype of black laziness seems to have been substantially internalized by African Americans themselves," the authors comment (Massey et al., 2003: 146–7).

[19] On the housing-related examples, see Anderson (2010: ch. 4). For the education ones, see, among many possibilities, Ferguson (2007). For the medical ones, see Institute of Medicine (2002).

I note that implicit bias research supports the internalization view, in that it finds that blacks have negative associations with blacks as a group, though not to the same extent as non-blacks (Jost, Banaji, and Nosek, 2004).[20] Why does Steele resist the commonsense view that if an ideology or stereotype is pervasive in a society, the group targeted in it will to some extent internalize it? He does not address this question explicitly. But one reason worthy of consideration in its own right and plausibly attributed to Steele is that he sees his accomplished student subjects as unlikely to have been plagued with crippling self-doubt, or they would not have become so accomplished.

This is a weak reason in that racial self-doubt does not have to be so overwhelmingly disabling. A person could have self-doubt yet possess other psychic resources for countering that self-doubt. I taught high school for a time and had several black students who expressed negative views of their race; several said that blacks were lazy or did not work hard. Sometimes my students were perplexed as to why (as they believed) blacks had these characteristics. They sometimes held these views of themselves and said they themselves did not work as hard as they should. But these views were far from entirely disabling to these students. By and large they did in fact work hard—some more than others—and so were obviously able to find motivational resources within themselves that countered the direction in which their racial self-doubts took them.[21]

A second reason Steele might reject internalized racial self-doubt is a concern about "blaming the victim." The idea that blacks subvert themselves by their own doubts is often taken, especially by conservatives, to mean that black underperformance is blacks' own fault. But the structural and social justice perspective that I have been arguing for here helps us to see what is wrong with that way of looking at internalized stereotyping. First, underperformance is in good part a product of a range of historical and structural features of the society, mentioned in Section 6, not primarily the psychology of black agents. Second, to the extent that psychological factors such as self-doubt do figure into the explanation of underperformance, they themselves are a product of the pervasiveness of anti-black stigma and stereotyping. It is a blinkered view of such psychological factors to fail to recognize them not as an autonomously self-produced, self-destructive

[20] Echoing an earlier point, I would caution that the mere unconscious negative associations that define implicit bias seem to me insufficiently robust to count for the possession of a stereotype or even a "negative view" of one's racial group.

[21] I mention these high school students because my black college students have not expressed these views to me so forthrightly, possibly because high school students are more open or less censored than college students as a rule. For a discussion of the issue of internalized racial self-demeaning (not only with respect to academic performance) among these high-school students, see Blum (2012: chs. 7 and 9).

culture among blacks, but as an expected product of segregation, inequality, and stigma (Anderson, 2010: 79).[22]

Thus deleterious images of blacks—including both the stereotypes that Steele and his colleagues point to as the source of stereotype threat, and the internalized versions of those that I am suggesting are an expected causal effect of their pervasive presence in society—cannot be fully eradicated from society unless society itself is made more racially equal. So the endeavor to rid society of anti-black stereotypes, stereotypes that cast doubt on black capabilities, has to be part and parcel of a struggle for a more racially just society. Although Steele would not deny this point in theory, his approach does not encourage recognizing it.

10 Conclusion

I have criticized the stereotype threat paradigm and research carried on within it on two general grounds. First, as an engagement with the world of education, while it helps deal with one piece of the undoubtedly important issue of racial disparities in achievement, it is quite narrow when looked at in the broader context of stereotypes and stereotyping generally. Stereotype threat research does not look into ways that students in stereotype-vulnerable groups can be equipped with the intellectual tools to recognize, examine, and criticize stereotypes in general and stereotypes of their group in particular. For example, it does not point the way to, nor hold out any hope for helping students of all groups to be able to distinguish stereotypes from valid generalizations. Nor does it encourage or draw on the concern members of such groups do or could feel for one another to enlist them in a collective project of doing battle with stereotypes that harm their group.

The second criticism is that the stereotype threat paradigm gives insufficient credence, or at least attention, to systemic injustices against stereotype-vulnerable groups. One part of this critique is that we need to recognize, and teach students, the difference between a stereotype and a valid generalization in part so that valid generalizations that are essential to recognizing, describing, and characterizing systemic racial injustices will not be avoided for fear of the taint that

[22] At one point Steele suggests a slightly different but related reason for rejecting internalized stereotyping. "If identity threat were rooted in an internal psychological trait, a vulnerability of some sort, then it would be harder to remedy. Would there be enough therapists to go around?" (Steele, 2010: 151). This is of course not a reason against the hypothesis itself but only one for hoping it is not true. In addition, as I have argued, we can acknowledge that self-doubt is not so disabling as to require only therapy as an intervention. It might also be that the adverting to therapy in this context reveals a failure to see internal vulnerability as intimately connected to unjust but changeable structures.

appropriately attaches to "stereotyping." I also argued that Steele's keeping at arm's length the idea that members of stereotype-vulnerable groups are prone to internalize negative stereotypes of their group is connected to his failure to recognize as a core part of his analysis the systemic and multifaceted nature of race-based injustice.

The individual learner-focused and the system-focused criticisms are connected. For in order for individuals to learn to avoid stereotyping themselves and, more generally, to see themselves as challenging racial stereotypes in society in general, they must recognize the embeddedness of stereotypes—as both cause and effect—in social structures and systems. That is, they must recognize that endeavoring to make the society itself more just is ultimately the only sure way to diminish stereotypes that rationalize injustice. Steele's way of talking about the stereotypes that cause stereotype threat almost takes the presence of those stereotypes as an unchangeable given. No attention is given to how we might undermine the presence of those stereotypes in the wider society, nor to how they might have been weakened historically.[23] Steele proposes as the chief educational task how to get individual members of stereotype-vulnerable groups "out from under" the stereotype, for example (as mentioned earlier) by reassuring the student that the salient figures in her academic environment do not believe the stereotype, or by reframing the situation so that the stereotype no longer applies to it. Steele does not ask what can be done about the stereotypes themselves, that is, about their presence in the culture of the society. The approach is not to get the subject to think "I recognize that stereotype and it is wrong and ignorant, and so I will not let it bother me—it's the stereotyper's problem, not mine"—but rather "that stereotype does not apply to me now, in what I am doing at the moment."

Finally, the individual critical thinking and the social-systematic criticisms are linked because it is plausible to think that a secure conviction that the disadvantages suffered by your own racial group, including educational disadvantages, are a product of unjust, discriminatory processes can be an important psychic resource for withstanding what you take to be others' questioning of your (and your group's) intellectual abilities. Steele and his colleagues do not, as far as I can tell, investigate the connection between vulnerability to stereotype threat and subjects' degree of conviction that the disadvantages their group suffers are due to unjust treatment (or a legacy in the present of such unjust treatment in the past). Nor do they investigate the related connection between how vulnerable an individual is to stereotype threat and how convinced she is that the view that

[23] Anderson shows how stereotypes and prejudices against blacks have changed in a positive direction in the past sixty or so years, yet still remain negative and potent (Anderson, 2010: ch. 3).

others hold of their group's intellectual deficiency is due to those others buying into dissonance-reducing and injustice-rationalizing ideologies of their own superiority, such as Anderson delineates. It seems to me at least worth exploring whether members of disadvantaged minorities who hold these convictions very firmly possess certain psychic resources that render them less vulnerable to stereotype threat than those who do not.[24]

If this "psychic resources" variable is indeed relevant to vulnerability to stereotype threat, it would provide a positive element in African American identity, an identity that is portrayed almost entirely in terms of achievement-undermining vulnerability in Steele's work. A community of African Americans can help its members acquire and sustain a sense of the group injustice to which they have been subject—to see their disadvantage not as a reflection of their deficiencies but as a moral deficiency in the society in which they live, and in whites who hold false ideologies that rationalize their own privilege. As African Americans have a history of resistance to racist treatment, there are communal and historical resources that African American students could draw on to sustain them and each other in taking up such a stance to stereotype threat—one that fully acknowledges that their white teachers and classmates may have stereotypic suspicions of their inferiority, yet does not allow these suspicions to undermine their own ability to believe in and muster their own intellectual resources.[25]

[24] Since Steele says that his subjects lack racial self-doubt, it might seem as if they implicitly hold to the conviction that the source of their group disadvantage is unjust treatment and false ideologies held by the majority and dominant group. But merely lacking *personal* racial self-doubt is not the same as affirmatively taking on the political perspective that provides a framework for seeing one's group's disadvantages as a product of unjust processes and false ideologies.

[25] I am influenced in these final remarks by having attended a conference on implicit bias in which Charles Lawrence—an African American law professor who wrote an influential article on unconscious bias—suggested to Greg Walton—a leading member of the stereotype threat research community—that a different way of thinking about stereotype threat that conformed to his own experience in law school when he was one of a small number of blacks, was not to try to get the subject out from under the stereotype. Rather he (and his peers) continued to recognize the presence of the stereotype among the whites at his law school but recognizing the pervasiveness of the stereotype was part and parcel of his recognizing the systemic injustice to which blacks were subject, one manifestation of which was the small number of blacks in the law school. This social structural approach allowed Lawrence to keep from being psychically undermined by the stereotype *without* being out from under it. He saw it as the false view of his white peers and teachers, as their problem as it were, as something that did not reflect on him. He had adopted a political identity as someone who was committed to struggling against systemic racial injustice and this helped keep stereotype threat from undermining his academic performance as an individual, while also giving him a civic and political perspective that provided a concern with more than how he himself was doing, to embrace a concern with his fellow blacks and with the social injustice to which they were subject.

I have not questioned Steele and his colleague's claims that there is such a thing as stereotype threat, that it contributes to academic achievement gaps between blacks and whites, and that interventions can reduce it in some education settings. Nevertheless, I have argued that the stereotype threat research paradigm contributes to depoliticizing a politically significant domain of educational disparity by (1) failing to situate stereotypes within social structures, (2) masking the influence of structural factors on educational disparities more generally, (3) failing to equip learners with civically significant tools for criticizing stereotypes of their groups, and stereotypes more generally, and (4) failing to see that in some ways academically vulnerable identities (such as "black" or "African-American") can play a positive role in warding off the effect of socially salient destructive stereotypes of their group.

References

Aronson, Joshua and Claude M. Steele (1995). "Stereotype threat and the intellectual test performance of African Americans." *Journal of Personality and Social Psychology* 69(5): 797–811.

Anderson, Elizabeth (2010). *The Imperative of Integration*. Princeton, NJ: Princeton University Press.

Blum, Lawrence (2004). "Stereotypes and stereotyping: A noral analysis." *Philosophical Papers* 33(3): 251–89.

Blum, Lawrence (2012). *High Schools, Race, and America's Future: What Students Can Teach Us About Morality, Diversity, and Community*. Cambridge, MA: Harvard Education Press.

Blum, Lawrence (2015). "Race and class categories and subcategories in educational thought and research." *Theory and Research in Education* 13(1): 87–104.

Bobo, L. and Michael Massagli (2001). "Stereotyping and urban inequality." In Alice O'Connor, Chris Tilly, and Lawrence Bobo (eds.), *Urban Inequality: Evidence from Four Cities*. New York, NY: Russell Sage: 89–162.

Duncan, Greg and Richard Murnane (2011). "Introduction." In G. Duncan and R. Murnane (eds.), *Whither Opportunity: Rising Inequality, Schools, and Children's Life Chances*. New York, NY: Russell Sage.

Farkas, George (2008). "How educational inequality develops." In Ann Chih Lin and David R. Harris (eds.), *The Colors of Poverty: Why Racial and Ethnic Disparities Persist*. New York, NY: Russell Sage.

Ferguson, Ronald (2007). *Toward Excellence with Equity: An Emerging Vision for Closing the Achievement Gap*. Cambridge, MA: Harvard Education Press.

Gates, Jr, Henry Louis and Claude Steele (2009). "A conversation with Claude M. Steele: Stereotype threat and black achievement." *Du Bois Review* 6(2): 251–71.

Gendler, Tamar (2008). "Alief and belief." *Journal of Philosophy* 105(10): 634–63.

Institute of Medicine (2002). *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care*. Washington, DC: The National Academies Press.

Jost, John T., Mahzarin Banaji, and Brian Nosek (2004). “A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo.” *Political Psychology* 25(6): 881–919.

Ladd, Helen (2011). “Education and poverty: Confronting the evidence.” *Journal of Policy Analysis and Management* 31(2): 203–27.

Loury, Glenn (2002). *The Anatomy of Racial Inequality*. Cambridge, MA: Harvard University Press.

Maass, Anne, and Mara Cadinu (2003). “Stereotype threat: When minority members underperform.” *European Review of Social Psychology* 14: 243–75.

Massey, Douglas, Camille Z. Charles, Garvey Lundy, and Mary Fischer (2003). *The Source of the River: The Social Origins of Freshmen at America’s Selective Colleges and Universities*. Princeton, NJ: Princeton University Press, 2003.

Grutter v. Bollinger, et al. 539 U.S. 306 (2003).

Oakes, Jeannie (2005). *Keeping Track: How Schools Structure Inequality*, 2nd edn. New Haven, CT: Yale University Press.

Oakes, P., S. A. Haslam, and J. C. Turner (1994). *Stereotyping and Social Reality*. Oxford: Blackwell.

Pollock, Mica (2005). *Colormute: Race Talk Dilemmas in an American School*. Princeton, NJ: Princeton University Press.

Rothstein, Richard (2013). “Why children from lower socioeconomic classes, on average, have lower academic achievement than middle-class children.” In P. Carter and K. Welner (eds.), *Closing the Opportunity Gap: What America Must Do to Give Every Child an Even Chance*. New York, NY: Oxford University Press.

Schofield, Janet Ward (1989). *Black and White in School: Trust, Tension, or Tolerance*. New York, NY: Teachers College Press.

Shelby, Tommie (2003). “Ideology, racism, and critical social theory.” *Philosophical Forum* 34: 664–92.

Shelby, Tommie (2011). “Justice and racial conciliation: Two visions.” *Deadalus* 140(1): 95–107.

Stangor, Charles (ed.) (2000). *Stereotypes and Prejudice*. Philadelphia, PA: Psychology Press.

Steele, Claude M. (2010). *Whistling Vivaldi: And Other Clues to How Stereotypes Affect Us*. New York, NY: W. W. Norton.

Steele, Claude and Joshua Aronson. (1998). “Stereotype threat and the test performance of academically successful African Americans.” In C. Jencks and M. Phillips (eds.), *The Black–White Test-Score Gap*. Washington, DC: Brookings Institution Press.

Stephan, Walter (1999). *Reducing Prejudice and Stereotyping in Schools*. New York, NY: Teachers College Press.

Tyson, Karolyn (2011). *Integration Interrupted: Tracking, Black Students, and Acting White After Brown*. New York, NY: Oxford University Press.

Wilson, William Julius (2001). *The Bridge over the Racial Divide: Rising Inequality and Coalition Politics*. Berkeley, CA: University of California Press.

2.2

Reducing Racial Bias

Attitudinal and Institutional Change

Anne Jacobson

1 Introduction

Many commonly available internet dictionaries define racism first in terms of beliefs. The resulting understanding of racism can easily make it seem that racism evaporates as the relevant beliefs change. Thus many Americans at various times have taken the United States to have entered a post-racial phase, while African Americans still had to suffer "the acts of everyday racism—remarks, glances, implied judgments—that flourish in an environment where more explicit acts of discrimination have been outlawed" (Chiasson, 2014). Such acts are a sign that racial bias still exists, even when the agents themselves would sincerely deny they are racist. How much does implicit bias account for the racism remaining in the United States? (See the Introduction in Brownstein and Saul, 2015, for an explanation of "implicit bias.") And how important is it to try to diminish such bias? In particular, will eliminating it finally create a just society?

In asking such questions, it is important to notice the institutional racism still present in the United States. The many dimensions over which African Americans are in the bottom ranks of society—medical care, education, wealth, salary, longevity, and so on—make it unlikely that a just society will be secured by a change in attitudes alone.

How then do we approach both changing individual attitudes and changing racist institutional elements? There have been serious challenges to the idea that we can simply tackle both together. In this chapter we will be concerned with issues surrounding such challenges.

We can start with reminders that in some cases reducing implicit bias may be relatively ineffectual or even irrelevant. As Dixon et al. (2012a) observe: "Kalev, Dobbin and Kelley's (2006) recent study of the shifting racial composition of 708 American organizations, for example, found that interventions to reduce managers' racial biases were comparatively ineffective as a means of implementing racial diversity. (A more effective strategy was to create institutional structures that delineated clear lines of responsibility and accountability for change in the workplace.)" The interesting point that discrimination may occur even without the engagement of bias against members of a group was highlighted in a recent Harvard Business Review (DiTomaso, 2014); what may be operating instead is a white network:

> But what if you're not a former associate of anyone in any employer that might need your skills? What if you're not a mentee, a friend of a friend, a relative? What if you don't come recommended by a trusted source—if you don't have an "in" of any kind and are not a known quantity? Then you're out of luck [in getting a job the usual way, through contacts], and that's exactly why today's corporate executives are missing the point about diversity: Whites don't have to do bad things to minority groups in order to maintain a racial advantage in employment and wealth. They only have to do good things for one another. And they do good things for one another all the time.

There is, then, evidence that our attention needs also to be directed to factors other than bias that are holding injustices in place, such as the institutions in a society that support such inequalities. Such a position can be discerned (albeit in different words) in earlier writers, such as Frantz Fanon (1965). Indeed, it is a common conclusion of post-colonialist theories. One central claim made is that the relationships made possible by a diminution of implicit bias may well not take us significantly closer to a just distribution of property and power, while it can distract the disadvantaged from pressing for it.

An important point, then, is that having oppressors take a less biased view of the oppressed can distract the oppressed from continuing to demand change even though the impact of the more positive views on the welfare of the disadvantaged may be slight. In the end we will propose a more complicated picture of how justice can be promoted than either the account proposing working on converting individuals or that pointing us toward transforming institutions (Shafir, 2013). These are two alternatives that may seem to point us toward an obvious third: taking up both causes. Nonetheless, combining the two original approaches may involve making serious compromises. Given that a reduction in bias may lead to a reduction in desire for change on the part of the oppressed, we will risk significantly slowing down social transformations when we simply combine the two. We will pick up this topic at the end of this chapter. As we will suggest, it will be hard to maintain just arrangements in a society that is indifferent to social justice.

In Sections 2 and 3 we will look at the role of both individual attitudes and features of institutions in creating and preserving significant racial injustices. We will start with the role of psychological attitudes as causes of biased actions. It is here that Dixon's appeal to disruptive interventions gets a solid grip.

2 Bias as Cause

There are reasons that could lead one to say that biases are not—or may not be—internal states that function as causes. This conclusion can make it difficult to even discuss the alternatives we have described. That is so because such a view could make it difficult to distinguish between individual traits and institutional features. Our approach to this problem will start with some recent cognitive science.

One of the most interesting claims of recent cognitive science and cognitive psychology is that we are very often wrong about the abilities we have (Chabris and Simons, 2010; Montague, 2007). For example, we think of vision as, when working rightly, enabling us to know the detailed facts of our environment. But, in comparison with our beliefs about it, vision is much more partial. As Pylyshyn noted some time ago (Pylyshyn, 2000):

> ...less information is encoded with each glance than has been previously assumed. Research by several workers has shown that information about the properties and relative locations of small changes in a scene are rarely noticed during saccades. Nevertheless, humans have the impression of a large, panoramic scene. Such a scene does indeed exist, but it is in the real world and not in the mind. (203)

And in addition, memory succeeds at giving us a usually useful gist of our environment, but not a detailed and accurate picture. Further, memories change in a fairly systematic way, whenever they are retrieved and then replaced by other memories.

One factor—which might also be considered a set of factors—that decides what we see is attention. As a famous experiment demonstrates, if we are concentrating on the passage of a ball among players, we may fail to see a large and active agent in gorilla costume start to frolic among the players (Chabris and Simons, 2010).

For a number of reasons, among them the incompleteness of vision that we have just mentioned, and the role of attention in it, there is now a considerable debate among philosophers about whether the psychological/visual experience is wholly within one's head (Noë, 2004; Prinz, 2006; Rowlands, 2003, 2010; Shapiro, 2011). The content of a report of someone's visual experience may far exceed the content of the inner experience itself, given that it has content.

Since such gaps may occur in many of the sensory experiences precipitating actions caused by bias, they may seem to threaten the idea of bias as an internal experience and cause, and, along with that, a distinction between individual and institutional bias. Assuming such "externalist" theories of visual perception are correct, then it may be wrong to think of some implicit biases as experiences internal to agents. Biases that seem to affect visual perception—for example, shooter bias (Correll et al., 2002)—may be conceived as institutional causes of behavior rather than as mental causes of perception. If correct, this would in turn threaten the distinction between individual and institutional bias.

At the same time, we can make a very strong case for saying that there are internal factors that we may be unaware of and that do cause discriminatory actions, such as unjustly grading disfavored students. That is, there are causally internal, individual, active implicit biases. The action of such biases is especially well investigated in Read Montague's labs, particularly that part done by Anne Harvey et al. (2010). (Though approaching the topic through negative biases and not positive ones, the work directed by M. R. Banaji in Phelps et al., 2000, is in close agreement with the neurophysiology.)

Harvey's work follows a decision making process that issues judgments about a kind of object, B, a painting, which has A, a corporate logo, on it. The subjects of the experiment are told that their participation is paid for by various corporations, and that these corporations contributed different amounts of money. As the case is set up, labels on the paintings remind observers about how much a sponsor paid for the experiment, where higher is better. It turns out that the subjects strongly prefer the paintings with high paying sponsors, though they—the subjects—are completely unaware of any influence.

FMRI investigations make us aware of what is going on inside of our brains when A's mere presence makes us judge B's quality differently. In the end stage, the activations of the dorsal lateral prefrontal lobes (DL-PFC) in effect create the link between evaluation and action. The ventral medial prefrontal cortex (VM-PFC) collects results from earlier activations that register rewards and disincentives. In the bias cases, the DL-PFC then provides the final activation. These internal factors then cause the biased action or judgment. While the content we ascribe to them may refer to many external factors, the causal action directly producing the actions and judgments is not employing these extra factors.

If we want to think about the moderation or elimination of bias, the next experiment from the same group gives us a lot of information. The researchers brought together a group of art experts and subjected them to the same picture-viewing paradigm. In the outcome in such cases, there was no sign of bias. The application of expertise at least in some cases appears to protect one against bias (Kirk, Harvey, and Montague, 2011).

Further fMRI studies provided a very revealing difference between the art experts and the naïve subjects. In particular, there were important differences in the activations of the brain regions responsible for registering rewards and those that provide the organism's final reaction to such rewards. One, the ventral medial prefrontal cortex (VM-PFC), in the naïve subjects was definitely affected by the pictures and their accompanying logos. The other, the dorsal lateral prefrontal cortex, underlies the final judgment on action, and in the naïve subjects there were very definite evaluative reactions.

The VM-PFC of art experts, on the other hand, was not affected by the difference in reward information that happened to accompany the pictures. It is not that the DL-PFC overcame or overrode the VM-PFC; rather, the biasing mechanism most of us have simply failed to deliver its standard result in experts. Interestingly enough, a few of the original naïve subjects did not give biased judgments about the art, and the VM-PFC was similarly less active as the art experts' were.

The research we have just looked at provides the following model: the VM-PFC collects assignments of values, and one outcome looks or feels best, or at least better than some others. The look or feeling may, however, not be conscious. Further, it may be just associative; that is, one may not have the goal of giving members of a racial group lower grades, or lower pay, but doing so results from an associative bond between social groups and evaluations. As Hume would say, it is instinct, not reason, that leads us (Hume, 2000). As we might put it, our judgment fits with our more instinctive "System 1" level of judgment (Kahneman, 2011). "Instinct" should be understood informally, merely making a contrast with explicit reasoning.

We can say that the concentration on the neural steps from assessment to action tells us how bias gets into action. We end up with a clear picture of how a bias is activated and produces the biased action. Such a picture is at least in accordance with a stress on the individual in the causing of bigotry's social effects.

A stimulus that gives us a very active VM-PFC in one person may not have the same result in another; something that frightens one person may leave another unmoved. Further, studying the differences between cases of the presence of implicit bias and its absence can give us some general ideas about how to overcome its effects. For example, in the picture viewing paradigm expertise about art renders conjunctions of pictures and monetary sums without effect on the VM-PFC. Knowledge can silence bias, as we hope to see in cases where the more objective merits of a case win out over personal preferences. It is, though, controversial whether even the Supreme Court always manages to overcome personal biases that fail to have the facts in their favor; knowledge does not always trump bias. It is also clear that there are ways to dampen the activity of the VM-PFC. Factors such as regulations with accompanying penalties for ignoring them may, for example, change the expected rewards that have led to much of the activity in the VM-PFC.

Given the work of neuroscientists we saw above, we have constructed a fairly specific idea of what it is to take implicit bias out of the causation of one's actions. When bias results in a biased action, a causal nexus in the frontal lobes enables a connection between bias and action. The research question many are asking is how to disrupt or diminish that connection. And there are a number of different answers, including imposing penalties, the expectation of which makes the biased option less attractive (Harvey et al., 2010; Kirk, Harvey, and Montague, 2011).

Action that we judge biased need not, however, always involve the mechanism we have just looked at. Such cases form another argument for institutional change as necessary. One kind of contrasting case consists in a very standard way in which human beings react to outsiders; that is, those not of their group. An outsider's presentation at a conference, for example, may even be judged meritorious, but insiders will tend to forget it and, unlike the work of insiders, it will not enter into the general discourse about what is going on in informal meetings after conference sessions (Burgess, Ryn, Dovidio, and Somnath Saha, 2007). In such a case the direct target of change appears to be attitudes, but it may be very difficult to change them without changing who will get counted as an outsider. To change the assignment of outsider status to members of various groups may be very important, and the way to effect such a change may require changing the institution's recruiting and admission standards, or criteria for sufficient experience, and so on.

Still another attitudinal source of bias can be found in a tendency many—and perhaps most—human beings have to greatly prefer fairly static institutional structures, particularly when the institutions are important to them. (It is an often repeated mantra in academic administration that faculty detest change.) Many people often oppose any change, while others will want to limit change greatly. Changes that get resisted include changing the languages allowed in a school or religious ceremony, changes that reorder or rework the "chain of command," changes in admissions standards for schools, professional organizations, informal clubs and so on. Changes that give a particular gender, race, or nationality access where it did not have access before may arouse great opposition.

The last two factors in attitude formation—about outsiders and about protecting institutions from change—also can obviously lessen the effectiveness of other attitude changes. In fact, these two problems can illustrate a limitation to simply changing attitudes. Attitudes may come in webs, and observable change may depend on acting on more than one point in the web. For example, the change in attitude may not be one that gives the outsiders the respect they would get in a just society. We saw this when US universities began to admit women to philosophy PhD programs, and some problems women encountered persist today, as entries in the blog "What is it like to be a woman in philosophy" strongly suggest.

"Intersectionality" is a label for the way in which discriminatory factors occurring together have causal powers beyond some simple addition of their impact. The problems faced by a disabled black man will likely not be simply those suffered by a white disabled person added to those suffered by a typical person of color, at least to the extent that new and important problems may arise. For example, a disabled man who is black may not receive any assistance if he falls while at a dangerous intersection, because people may see African Americans as threatening. People of color, then, may have a much more difficult time when disabled.

Before we move on to consider institutions, we should consider how a stress on bias and attitudes may be viewed from a more ideological perspective. Some have suggested that focusing on individuals leaves out important elements of community (Bird, 1999; Peters and Marshall, 1996; Royce, 2009). Famously, Descartes articulates an individualistic theme when he proceeds to attempt to discover his own essence and to review what can be understood; he does this with the hypothesis that he is entirely alone. We can see this as individualism taken to an extreme.

The Cartesian stance contrasts markedly with the explanatory formats more suited to collectivist cultures. Collectivist views theorize individuals in terms of the groups in which they are found; individualist theories see, rather, the group as explained by the individuals. The stress on individuals' attitudes in this chapter's opening paragraph as the cause of injustice in the society at least echoes the individualistic approach. The impact of these different cultures and explanations may be very profound. Neuro-imagining studies strongly suggest that people from a collectivist society in effect include other people in what we might call their self-concept (Chiao et al., 2009; Zhang et al., 2006). Westerners typically do not.

One emerging criticism, then, resides in the fact that collectivist cultures do not see the individual's attitudes independently of the culture, and this may limit our ability to transfer an effective strategy against bias from one culture to another.

3 Focusing the Question: Individuals or Institutions or Both?

Despite some unresolved problems, Harvey and Montague's work gives us half the picture we need to discuss the current stress on eliminating or mitigating implicit bias. But to determine whether altering the brain's actions and reactions will or can take us far along to the just society we want, we need to look more closely at the context in which the bias operates. The latter may well require actions quite different from, or additional to, mitigating or eliminating implicit bias. If so, then we need to consider the possibility that changing attitudes may not do much, or at least not enough.

The contrasting approach to bias that Dixon gives us distinguishes between moderating attitudes and attacking institutions (Dixon, Levine, Reicher, and Durrheim, 2012a, 2012b). We can understand changing institutions as involving changing more than the attitudes of individuals. For example, diversity initiatives in institutions may aim at changing the racial or sexual proportions of employees by changing racist and sexist attitudes, but if they only address attitudes they will not in our terms be considered as operating on the institution. In contrast, introducing a rule that requires that each syllabus contains an author from an underrepresented group is an institutional change, albeit a small one.

We have seen the distinction between individuals and institutions as factors in a number of Western countries as the debate about same-sex marriage has occurred. Getting to the point where same-sex marriage is feasible has required, at least in practice, changing both attitudes and institutions, such as the laws governing marriage. We can perhaps too easily imagine the citizenry of a country being 90% in favor of allowing same-sex marriage, while their very conservative legislature refuses to enact the needed legislation. In such a case, we may find that same sex couples still cannot be considered married. This is a case in which changing attitudes has not been enough to create a more just situation. At the same time, it is a clear case of attitude change leading to an institutional change.

In asking whether changing individuals' attitudes is enough to create a just society, we do not need to require some sort of instantaneous change for an affirmative answer, as though we could change attitudes on Monday and find we have a just society on Tuesday. Rather, what we want to know is whether further significant changes will be required. An analogy: it is not enough to provide flowers for a friend in a hospital to choose the flowers to be displayed. Rather, there are many other issues that need to be addressed and resolved, from getting the flowers to the right place in the hospital onto finding the right sort of vase, which may involve a change in what is in the supply cupboard. The vase analogy has a serious analog that health care workers, among others, complain about, illustrated by the US donation of drug supplies to countries which lack the means to distribute the drugs, as Elizabeth Reid, former director of the UNDP program for AIDS in developing countries, has emphasized in private conversations. Here the institutional changes may also include introducing institutions, such as rural health clinics.

It is useful to categorize different kinds of areas of social change in order to understand whether attitude change will be enough to solve institutional injustices in the United States as it is today. Health, wealth, housing, and education give us a promising beginning, though we are leaving out such important areas as law (Levinson and Smith, 2012). We need to look at how the selected areas intersect in creating failures of justice. For example, implicit bias looks to be, given the literature, a very significant problem in medical care (Burgess et al., 2007; White and Chanoff, 2011). (White recounts the experience of a young resident in asking a very distinguished cardiac surgeon why only his white male patients received angioplasty; the aghast surgeon had been completely unaware of his selectivity.) However, even if we reach a state where, for example, whites and blacks asking for medical care receive the same treatment for heart problems and pain relief, to pick two notorious areas of unequal treatment, the problem of health insurance can come close to making health proportional to wealth; in states that have refused additional federal funding for Medicaid, "Obama Care" still leaves millions without health insurance. Thus the drastically lower wealth in the African American community means that inequality will survive attitude change, since money is required for most things.

The policy of funding medical care through health insurance has not been thought up to target black people directly, and neither have a number of the other factors we will look at. Nonetheless, the way such factors affect black people contributes to their markedly poorer medical treatment.

Let us start, then, with health. The CDC claims: "Health disparities between African Americans and other racial and ethnic populations are striking and apparent in life expectancy, death rates, infant mortality, and other measures of health status and risk conditions and behaviors" (Control, 2013). When we look at studies of medical practice in the US, we see plentiful and detailed descriptions of the biased attitudes of many medical personnel (White and Chanoff, 2011). Changing them may change other factors for the good. For example, we may change the cultural barriers that reflect the negative and too often ineffectual and even hostile treatment African Americans can receive.

At the same time, it is also the case that other factors will interact; children from very poor families may be less well educated, in addition to receiving less good medical care. These factors are connected. We can expect that an inferior education leaves them, in addition to their parents, with reduced means to make healthy choices. In addition, poor, medically neglected children are going to have a harder time performing well in school. As ill-educated and poor, they will be less able to better their share of social goods.

We should be beginning to get the sense of a number of factors that together intersect to create a tangle of problems which discriminatory attitudes, whether explicit or implicit, are in part keeping in place. (Anderson skillfully links them with segregation; Anderson, 2010.) And confirmation of this interrelatedness can be found in the CDC's list of revealing disparities relevant to health (Control, 2013):

> African American women and men 45–74 years of age in 2006 had the largest death rates from heart disease and stroke compared with the same age women and men of other racial and ethnic populations.
>
> Adolescent and adult African Americans ages 15–59 years in 2007 had the largest death rates from homicide, as compared with other racial and ethnic populations of the same ages.
>
> Hispanic American and African American adults aged 18–64 years had substantially larger percentages of uninsured populations compared with Asian/Pacific Islander and white Americans.

The first figure is arguably a product of, to some extent, medical bias. If you recall our earlier tale, a doctor may limit his or her best treatment to white men (without realizing it). But there are substantial institutional factors at work also, and so efforts in other directions are needed. One is the sheer cost of health care in the United States—in part because of health insurance, but also because of other factors, such as the pricing of medications. It is not so unusual to have important medications cost $10 a day after insurance has been applied. Drugs for cancer may be many times more expensive, and one's health insurance can refuse to cover the cost even if one's doctor says they are essential. In addition, the CDC lists:

> In 2009, high-school completion among African American adults was the second lowest (second to completion among Hispanic adults and similar to the completion among American Indian/Alaska Native adults).
>
> In 2009, the percentage of African American adults living in poverty was among the largest compared with other racial/ethnic populations (similar to percentages among American Indians/Alaska Natives, and Hispanic Americans).
>
> In 2009, African American adults more often lived in inadequate and unhealthy housing than white adults.

In the list of comparisons from the CDC, we can discern what institutional factors are operating. Thus, the education of an African American is not affected adversely simply because of the bigoted attitudes of their teacher. Among other things, the wealth of the parents will impact the schooling of the children. For example, place of residence contributes to how well funded the children's schools are, since a large part of school funding depends on local tax resources, though this is in the process of changing somewhat. Nor will poor parents easily find good alternatives to public schools. Hence, wealth will impact both health and education. There is good evidence that bigotry can affect the chances for finding good jobs. But more than a lack of bigotry is needed to qualify for almost all well-paying jobs.

Two important facts are becoming clear. One is that there are institutional factors creating or holding in place inequalities that require more than changing attitudes to change. The second is that factors intersect and make it difficult to find an effective starting point. Increased salaries may be easier if there is better schooling, but better schooling may require more wealth in the community.

Dixon challenges the idea that the route to social justice must include diminishing or eliminating implicit bias, or at least the ability of bias to shape actions (Dixon et al., 2012a, 2012b). Dixon's argument has two major components. First of all, he argues that working on implicit bias (or, in his controversial terms, helping people feel better about some members of a group) is not proving effective in producing more just situations. Secondly, he maintains that group agitation and action can be, and have been, effective in creating desired social change. Each part of his discussion is informative.

According to Dixon, there are a number of factors we should be aware of when we consider trying to address individual bias. One of the most important is that some techniques that have gotten a lot of attention will likely misfire in a way that the advantaged may well not see. That is, the improved attitudes of the advantaged may remain deeply affected by the stereotypes in the biased society. Thus, for example, someone may become keen to allow women into philosophical programs, but still retain a notion of women's philosophy and a belief that that is something all women can teach. Another problem lies in the fact that a partial reduction in felt bias may reduce the oppressed's desire for better solutions. For example, the sincere civility and care from medical personnel may well decrease the concern about the inferiority of procedures available to minorities or, too often, women.

For Dixon, the better approach to securing justice lies in group protest and agitation. In light of the previous discussion, we can divide his claim here into two parts: (1) institutional change may be necessary to secure the just results we seek; (2) such change is best accomplished by group agitation and protests. We have seen a lot of reason to be sympathetic to the first claim. Our look at the situations of African Americans in the United States reveals a number of factors that cannot be changed just by changing attitudes. The second sort of claim, though not unique to Dixon, has been challenged on its view of the facts.

To the extent that Dixon is right—as he seems to be—that recently the primary focus of bias research has largely left out the role of institutions, he is making a very important point. Furthermore, reflection on his arguments suggests another one: if we do not change the institutions, then change in attitudes may be relatively temporary. It is in many cases the effects of institutions that are at least partial causes of bias. For example, doctors are reluctant to treat African Americans, or to treat them very aggressively, because their rate of compliance with medication is very low (White and Chanoff, 2011). But it seems reasonable to suppose that some of the problems with compliance can have to do with the very high expense of many medications in the United States, and the very perplexing labyrinth of rules constituting Medicaid, which is for those with low income.

Here is another example, suggested by Sally Haslanger (2012): the lack of support for women with children keeps them too often from working well at demanding jobs. Even if all the gatekeepers that operate between women and demanding jobs were to change their attitudes, without structural changes the women will be less capable of matching men's performance, especially if the men do not have much responsibility for home and childcare. We should expect negative opinions about women workers to be given a second life in such situations.

Dixon's second claim is that protest is the best way to secure justice. The Civil Rights Act of 1964 has provided grounding for a number of changes that have reduced the amount of discrimination many US citizens would otherwise face. But arguably, as Dixon claims, this is a good example of agitation and protest producing more just opportunities for underrepresented groups. What should we say, then, of the recent number of states legalizing same-sex marriage? No doubt we should count the agitations of the 1960s as challenging the institutional setting for discrimination against gays, as a number of researchers have argued. As Poinddexter maintains:

> During the 1960s, several political and social upheavals heavily influenced each other. A philosophy counter to the mainstream culture emerged that supported equality, freedom, choice, and peace. The homophiles of the previous decade had inadvertently set the stage for members of the US gay community to consider their place in the societal chaos of the social protest era. The social context of choosing to be who you wanted to be had been altered. At the same time, the legal situation began to change, and there was

> some movement toward the decriminalization of the private, consensual sexual acts of adults . . . (Festle, 2005)

Nonetheless, in our current setting, attitude change played a significant role in the acceptability of same-sex marriage.

What this and other cases we have looked at suggest are two things: (1) whether we are dealing with an attitude problem or an institutional one will vary from one context to another, and we may not always be able to sharply distinguish among them, and (2) we must remember that institutions can deeply affect attitudes, so we neglect them at our peril. Even for those whose biased attitudes have changed to positive ones, gender or skin color may leave someone an outsider.

At the same time, if we neglect attitudes, bias may lead people to reinvent unjust institutions or resist the changes in old ones, among other things. It is also true that mitigating bias in the powerful may decrease salutary demands for change. In combining the two approaches, addressing both individuals and institutions, we have to compromise. Progress may well depend on there being enough people in a society who care about justice so that we do not find ourselves too often unable to reach beyond the status quo.

Acknowledgments

I am very grateful to Michael Brownstein and Jennifer Saul, whose very careful and insightful comments on two earlier drafts of this chapter were exceptionally helpful. Earlier work with William Langley was very helpful, as were comments by anonymous referees.

References

Anderson, E. (2010). *The Imperative of Integration*. Princeton, NJ: Princeton University Press.

Bird, C. (1999). *The Myth of Liberal Individualism*. Cambridge and New York: Cambridge University Press.

Brownstein, M. and Saul, J. (2015). "Introduction." In Brownstein, M. and Saul, J. (eds.), *Implicit Bias and Philosophy: Volume I*. Oxford: Oxford University Press.

Burgess, D., Ryn, M. v., Dovidio, J., and Saha, S. (2007). "Reducing racial bias among health care providers: Lessons from social-cognitive psychology." *Journal of General Internal Medicine* 22: 682–887.

Chabris, C. F. and Simons, D. J. (2010). *The Invisible Gorilla: Thinking Clearly in a World of Illusions*. New York, NY: Crown Publishers.

Chiao, J. Y. et al. (2009). "Neural basis of individualistic and collectivistic views of self." *Human Brain Mapping* 30(9): 2813–20. doi: 10.1002/hbm.20707.

Chiasson, D. (2014). “Color codes: A poet examines race in America.” *The New Yorker*, October 27.

Control, C. f. D. (2013). “Minority health.” <http://www.cdc.gov/minorityhealth/observances/BAA.html>.

Correll, J., Park, B., Judd, C. M., and Wittenbrink, B. (2002). “The police officer’s dilemma: Using ethnicity to disambiguate potentially threatening individuals.” *Journal of Personality and Social Psychology* 83(6): 1314–29.

DiTomaso, N. (2014). “White people do good things for one another, and that’s bad for hiring. *Harvard Business Review* blog.<http://blogs.hbr.org/2014/01/white-people-do-good-things-for-one-another-and-thats-bad-for-hiring/>.

Dixon, J., Levine, M., Reicher, S., and Durrheim, K. (2012a). “Beyond prejudice: Are negative evaluations the problem and is getting us to like one another more the solution? *Behavioral and Brain Sciences* 35(6): 411–66. doi: 10.1017/s0140525x11002214

Dixon, J., Levine, M., Reicher, S., and Durrheim, K. (2012b). “Beyond prejudice: Relational inequality, collective action, and social change revisited.” *Behavioral and Brain Sciences* 35(6): 451–9. doi: 10.1017/s0140525x12001550.

Fanon, F. (1965). *The Wretched of the Earth*. New York, NY: Grove Press.

Festle, M. J. (2005). Listening to the Civil Rights Movement (Vol. 12, pp. 10–15).

Harvey, A. K. U., Denfield, G. H., and Montague, P. R. (2010). “Monetary favors and their influence on neural responses and revealed preference.” *Journal of Neuroscience* 30(28): 9597–602.

Haslanger, S. A. (2012). *Resisting Reality: Social Construction and Social Critique*. New York, NY: Oxford University Press.

Hume, D. (2000). *An Enquiry Concerning Human Understanding: A Critical Edition*. Oxford and New York: Oxford University Press.

Kahneman, D. (2011). *Thinking, Fast and Slow*, 1st edn. New York, NY: Farrar, Straus and Giroux.

Kalev, A., Dobbin, F., and Kelly, E. (2006). “Best practices or best guesses? Assessing the efficacy of corporate affirmative action and diversity policies.” *American Sociological Review* 71(4): 589–617. doi: 10.1177/000312240607100404.

Kirk, U., Harvey, A., and Montague, P. (2011). “Domain expertise insulates against judgment bias by monetary favors through a modulation of ventromedial prefrontal cortex.” *Proceedings of the National Academcy of Sciences* 108(25): 10332–6.

Levinson, J. D. and Smith, R. J. (2012). *Implicit Racial Bias Across the Law*. Cambridge and New York: Cambridge University Press.

Montague, R. (2007). *Your Brain is (Almost) Perfect: How we Make Decisions*. New York, NY: Penguin Group.

Noë, A. (2004). *Action in Perception*. Cambridge, MA: MIT Press.

Peters, M. and Marshall, J. (1996). *Individualism and Community: Education and Social Policy in the Postmodern Condition*. London and Washington, DC: Falmer Press.

Phelps, E. A., O’Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., and Banaji, M. R. (2000). “Performance on indirect measures of race evaluation predicts amygdala activation.” *Journal of Cognitive Neuroscience* 12(5): 729–38.

Prinz, J. (2006). “Putting the brakes on enactive perception.” *Psyche* 12: 1–19.

Pylyshyn, Z. W. (2000). “Situating vision in the world.” *Trends in Cognitive Sciences* 4(5): 197.

Rowlands, M. (2003). *Externalism: Putting Mind and World Back Together Again*. Chesham: Acumen.

Rowlands, M. (2010). *The New Science of the Mind: From Extended Mind to Embodied Phenomenology*. Cambridge, MA: MIT Press.

Royce, E. C. (2009). *Poverty and Power: The Problem of Structural Inequality*. Lanham, MD: Rowman and Littlefield.

Shafir, E. (2013). *The Behavioral Foundations of Public Policy*. Princeton, NJ: Princeton University Press.

Shapiro, L. A. (2011). *Embodied Cognition*. London and New York, NY: Routledge.

White, A. A. and Chanoff, D. (2011). *Seeing Patients: Unconscious Bias in Health Care*. Cambridge, MA: Harvard University Press.

Zhang, L., Zhou, T., Zhang, J., Liu, Z., Fan, J., Zhu, Y. (2006). “In search of the Chinese self: an fMRI study.” Science in China: Series C Life Sciences 49(1): 89–96.

PART 3

The Ethics of Implicit Bias: Theory and Practice

3.1

A Virtue Ethics Response to Implicit Bias

Clea F. Rees

Virtue ethics faces two challenges based on psychologists' work on the role of automatic processes in cognition. Both arise from our reliance on cognitive processing which is relatively immune to direct deliberative control and of which we are relatively unaware. These challenges threaten not only the very possibility of virtue but, more fundamentally, our conception of ourselves as rational persons (e.g. Doris, 2009).

The first is the classic situationist challenge which suggests that much of our behaviour is determined by trivial and arbitrary features of situations of which we are unaware and which we would not endorse as reasons for action (e.g. Merritt, Doris, and Harman, 2010). Virtuous action is action done for the right reasons; behaviour cannot be virtuous if it is not motivated by reasons at all. Since virtue ethics is intended to guide the moral lives of creatures like us, this first challenge threatens not only the possibility of our realizing virtue, but the appeal of virtue ethics qua ethical theory.

One of the most promising responses to this first challenge argues that the influence of automatic processes on cognition facilitates, rather than threatens, rational agency and virtuous action because habituation can ensure that automatized cognition embodies the right motivations (e.g. Snow, 2009; Rees and Webber, 2014).

The second challenge threatens this response by suggesting that the influence of virtuous automaticity on cognition will be systematically undermined by our unwitting habituation of the wrong motivations. Virtue requires not only the habituation of virtuous motivations but the non-habituation or dehabituation of vicious ones which would otherwise undermine the connection between virtuous motivation and virtuous action. Implicit bias is an especially stark illustration of this second challenge: research suggests that we may be oblivious to the existence and behavioural influence of disturbing features of *ourselves* in the form of habituated associative biases which we have explicit reasons to reject (see the Introduction in Volume I). While it might be disconcerting to discover a disproportionate number of aspiring **Phil**osophers named **Phil**ippa and **Phil**lip (Pelham, Mirenberg, and Jones, 2002), that our implicit sexism might frustrate the aspirations of the former is positively disturbing. Moreover, whereas more troubling situationist results such as Milgram's (2009) depended on carefully engineered experimental manipulations (Russell, 2011), the threat to virtue posed by implicit bias requires only the reality of social prejudice. Although virtue ethicists recognize the crucial role of social support in developing and sustaining virtue, because virtue ethics purports to provide practical guidance, the pervasiveness of implicit bias rules out simply dismissing it as the product of a bad environment. Given that implicit bias occurs not only outside conscious awareness, but despite deliberative abhorrence, what counsel can the virtue ethicist possibly offer the implicitly biased?

Responses to the situationist challenge which appeal to virtuous habituation invoke a model of cognitive processing based on two areas of psychological research. The first is social psychologists' work on attitudes and attitude change in the context of an associative model of personality. The second is research concerning the automatization of goals.

In this chapter I argue that this model also offers the virtue ethicist a promising response to the second challenge, because automatization has the potential not only to habituate virtuous motivations, but to dehabituate vicious ones. In particular, the virtue ethicist can respond to the implicitly biased by counselling the habituation of egalitarian virtue, rather than merely the control of anti-egalitarian vice. Specifically, I argue that the habituation of individual egalitarian commitments is crucial to strategies of active resistance, and that communities should ensure the collective support this process requires.

Section 1 outlines dual-process models of cognition and the particular role of those which posit distinct systems for automatic and deliberative processing in accounts of the threat posed to virtue ethics by implicit bias. Section 2 explains why indirect mitigation strategies offer virtue ethicists an unsatisfying response to implicit bias given such models of cognition. Section 3 sketches the psychological structure of attitudes, as understood by social psychologists, and their role in an alternative model of cognitive processing. Section 4 explores the process by which consciously selected goals and commitments may be automatized, and explains its particular interest. Section 5 argues that a satisfactory defence of virtue ethics is supported by research on the automatization of strong egalitarian commitments. Section 6 explores some puzzling questions raised by current research and indicates how future work might seek to address them. Section 7 explores the potential of egalitarian commitments to alter our implicit biases themselves, and explains some further limitations of current research. Section 8 explains an important implication of my argument for virtue ethics: individuals can effectively habituate egalitarian virtue only if their communities share their commitment to resisting implicit bias.

1 'Dual-Process' Models and Implicit Bias

Psychologists have developed several 'dual-process' theories of cognition (Maio and Haddock, 2010: 96–106). For the purposes of this chapter, what is important about such models is their common claim that deliberative and automatic cognition are distinct kinds of cognitive processing. Deliberation is explicit, conscious processing which analyses information carefully and logically, is sensitive to the content and strength of arguments, and is relatively slow and effortful. In contrast, automatic processing depends on heuristics and learnt associations, is more sensitive to the source and form of arguments, and is relatively fast and effortless. Whereas deliberation might lead you to choose unbranded paint for your home (because it was cheaper) or branded (because it was higher quality), associative processing might result in a choice of Dulux (because you liked the dog in their advertising). Whereas cost and quality comparisons require conscious attention, you might be unaware of the canine influence on your décor.

Humans could not make do with only deliberative processing; automatic cognition is essential. Moreover, bias in the broadest sense provides crucial filtering enabling us to focus limited cognitive resources on what is of greatest importance to us. At its best, implicit bias attunes parents to the particular cries and needs of their own children, allows surgeons to focus on critical features of the body in front of them, and enables examiners to assign marks informed by the features of essays of greatest disciplinary relevance. In the complete absence of such bias, every mother in the maternity ward would need to consider every cry in order to decide whether to respond to it, the surgeon would require attentional resources to ignore your appendix when removing your tonsils, and examiners would need to consciously set aside students' choice of ink as irrelevant to their knowledge of Kant's metaphysics.

Our capacity for conscious, effortful cognition is limited. Attention focused on one task cannot be simultaneously devoted to others, and the expenditure of volitional resources affects their subsequent availability (Baumeister et al., 1998;

Muraven, Tice, and Baumeister, 1998; Muraven and Baumeister, 2000). Time is another limited resource. Even if the dog is irrelevant to the quality and value of Dulux paint, the canine association may facilitate a perfectly rational choice. Although appealing to the dog would undermine the rationality of a deliberative decision to buy Dulux, his appeal need not undermine the rationality of a less considered choice if the costs of more careful deliberation would outweigh the benefits of a more considered one. It can be quite irrational to expend the resources required to reach a more rational decision. Moreover, we are constrained not only by the total time available to us for all tasks. but by the time-sensitive nature of many decisions.

In general, then, it is no bad thing that we rely on associative processing and heuristic short-cuts. Unlike our relatively innocuous paint purchases, however, other learnt associations are far from harmless. 'Implicit bias' in the problematic sense refers to biases we soak up from our social environment in the form of implicit morally problematic associations with characteristics such as race, sex, and sexual orientation (see the Introduction in Volume I). These problematic implicit biases influence cognition in ways which systematically disfavour members of non-dominant groups. Because such biases are systematic rather than arbitrary, the collective impact of individuals' implicit biases on members of non-dominant groups is likely to constitute a significant harm even when the impact of each instance is negligible (Brennan, 2009). Moreover, some instances will themselves constitute significant harm. If simulated decision-making provides a reasonable indication of its effects, implicit bias reduces the chances one will be interviewed for a job if one is Arabic rather than Swedish or hired to a managerial post if one is female rather than male, and makes it more likely that one will be shot by an armed police officer ('shooter bias') and less likely that one will receive appropriate treatment for coronary heart disease if one is black rather than white (Jost et al., 2009).

The worry is that because our implicit biases are acquired and utilized outside conscious awareness, we cannot perceive or correct for their effects. Dual-process theories which explain automatic and deliberative processing by invoking distinct, relatively independent systems of cognition deepen this concern by suggesting that even educating ourselves about the problem might leave us unable to alter or control our implicit biases.

However depressing this evidence might be from the perspective of policy makers and concerned citizens, one might think it not altogether bad news for virtue ethicists. After all, the evidence for their effects on decision-making depends on variation in the kind and degree of individuals' implicit biases. Just as one response to the situationist challenge emphasizes the rarity of virtue (e.g. Kamtekar, 2004), one might argue that the prevalence of implicit bias is simply further evidence of widespread ethical deficiency.

This response is ruled out, however, by a key tenet of virtue ethics: virtue can be developed. It is true that the right sort of early education may be essential: one may not be blameworthy for one's lack of virtue if one was deprived of appropriate habituation as a child. This might be mere common sense except for the failure of much moral philosophy to acknowledge it. That virtue is as much a collective responsibility as an individual one, and that the development of moral agency requires an appropriately supportive social context, will come as no surprise to feminist philosophers, educators, and parents (e.g. Baier, 1995).

In the case of implicit bias, however, it is not at all clear what the 'right sort' of education might be. Given that implicit biases are found even in individuals engaged in efforts to actively resist prejudice, it is unclear not only how such individuals should respond to their own implicit bias, but also what they might do to reduce it in the next generation. Implicit bias is not limited to unfortunates attempting to overcome the effects of explicit encouragement to prejudice.

Moreover, most virtue ethicists argue that even a poor start can be mitigated or overcome by later efforts. Scrooge's decision to reform begins a process of rehabituation which replaces miserliness and meanness with generosity and compassion (Annas, 2011:12). Partly because Scrooge is a self-conscious miser who despises kindness in others, reflective deliberation can instigate and guide self-reform. In contrast, given that implicit bias can occur not only outside conscious awareness but despite deliberative abhorrence, what counsel can the virtue ethicist possibly offer the implicitly biased?

2 Stocking the Egalitarian's Toolbox

Indirect mitigation strategies have proven effective in combating the behavioural effects of implicit bias. The virtue ethicist might therefore recommend that individuals and institutions respond by implementing these strategies themselves, raising awareness, and encouraging others to follow their lead.

First, institutions can select from a range of mitigation strategies. For example, the representation of female musicians in top orchestras improved partly due to the introduction of screens rendering candidates audible but invisible during auditions (Goldin and Rouse, 2000). Similarly, there is some evidence for a reduction in gender bias on referees' judgements when journals implement double-blind reviewing (Peters and Ceci, 1982; Budden et al., 2008; but cf. Blank, 1991). Ensuring that decision-makers anticipate needing to justify their decisions to an audience whose views they cannot predict can encourage more thorough scrutiny of the relevant considerations, more careful analysis of the pros and cons of various options, and reduced reliance on the automatized associations which constitute implicit bias (Lerner and Tetlock, 1999: 256–8, 263). Even when the views of the anticipated audience are known, accountability may be effective in inducing more careful analysis if deliberators are motivated to base their decisions on accurate evidential evaluations (Quinn and Schlenker, 2002). Although the conditions under which accountability is effective matter in reducing the effects of implicit bias on decision-making (Lerner and Tetlock, 1999: 258–9, 264–6), this need not undermine its effectiveness in a range of key cases. For example, while monitoring perceived as illegitimate can actually increase the effects of bias, the legitimacy of a requirement to justify personnel or prosecutorial decisions is unlikely to be doubted.

Second, in addition to encouraging and implementing appropriate institutional practices, a number of mitigation strategies are available to individuals. Envisaging or imagining counterstereotypic exemplars, or thinking oneself into others' shoes can help to overcome the effects of implicit bias on cognitive processing (Corcoran, Hundhammer, and Mussweiler, 2009; Dasgupta and Greenwald, 2001; Dasgupta and Asgari, 2004; Blair, Ma, and Lenton, 2001; Galinsky and Moskowitz, 2000). Forming 'implementation intentions' is another way for individuals to neutralize the influence of their implicit biases on behaviour (Webb, Sheeran, and Pepper, 2012). An implementation intention is a specific behavioural plan as opposed to a more general goal. 'I will study harder' is a general commitment; 'If it is 3 o'clock on a Tuesday, I will study in the library!' is an implementation intention.

Strategies which allow us to indirectly mitigate the effects of bias on our treatment and judgements of others have attracted considerable attention. Theoretical work in philosophy has recommended considering the availability and likely effectiveness of these strategies when choosing between competing normative ideals concerning racial categorization (e.g. Kelly, Machery, and Mallon, 2010), and leveraging them to satisfy epistemic and moral demands (e.g. Merritt, 2009; Kelly and Roedder, 2008). Saul (2012) has argued they should inform the philosophy Research Excellence Framework (REF) in the UK and the Philosophical Gourmet Report in the US. Furthermore, institutions have begun to actively promote their use. The US National Center for State Courts has produced educational materials encouraging their use to address the effects of implicit bias on judicial decision-making (Casey et al., 2012). In the UK, the Chair of the REF Philosophy Panel has responded to Saul's (2012: 263–4) concerns by ensuring that members are aware of the literature on implicit bias and of ways to reduce its impact, the Equality Challenge Unit (2013) is developing strategies to counteract its influence on recruitment decisions in higher education institutions, and Remploy (2013) offers practical ways to mitigate its effects on individuals with facial disfigurements.

There are good reasons for these recommendations: indirect mitigation strategies are crucial if only because no momentary act of will can eradicate implicit bias. Using such strategies enables us to mitigate the behavioural effects of our biases in especially sensitive or significant situations, especially ones of which we are aware and for which we can prepare in advance. Considered in isolation, however, the solution which such strategies promise seems neither psychologically nor theoretically satisfying, because it apparently offers us little hope of changing our implicit biases themselves. If automatized and deliberative processes involve distinct cognitive systems, then there is no obvious way for strategies which rely on deliberative control to alter the implicit associations whose influence they mitigate.

First, committed egalitarians concerned about their characters are unlikely to find the solution psychologically satisfying. Although I would much prefer that envisaging counterstereotypical exemplars prevent my biased associations from leading me to treat a short, blind, black, female resident of Merthyr Tydfil less well than a tall, able-bodied, white, male inhabitant of Ascot, I would prefer to alter my underlying bias itself. Indeed, I would wish to be free from implicit bias even if its behavioural consequences were entirely benign. Moreover, the effectiveness of indirect mitigation strategies is limited by our epistemic and cognitive capacities. Situations arise for which nobody can be fully prepared and the number of implementation intentions one can usefully form is presumably limited.

Second, while she should surely encourage their use by both institutions and individuals, indirect mitigation strategies appear to offer the virtue ethicist at most rather cold theoretical comfort. It is not sufficient for virtue that one reflectively endorse the right values and that one's behaviour reflect those values. What matters also is that one's habituated, automatized motivations embody them. This is the grain of truth in the myth that true virtue is effortless: alleviating another's distress may require considerable effort, but being moved to do so should not. A need to rely on indirect mitigation strategies to control the effects of one's implicit biases shows that one cannot rely on one's habituated, automatized motivations and thus reflects a deficiency in virtue. Dependence on such strategies shows that vicious automaticity would otherwise interfere with the influence of virtuous automaticity on cognition. Indirect mitigation strategies cannot restore virtue if they are limited to controlling, rather than eliminating, vicious motivation. Unless the virtue ethicist can say something about how control can be habituated and implicit bias itself reduced or eliminated, she is limited to counselling the control of vice rather than the development of virtue.

Fortunately, current psychological research offers a better defence of virtue ethics based on an alternative to dual-process models which posit distinct systems for automatic and deliberative of cognition. While the existing literature does not guarantee the viability of virtue ethics, it does provide grounds for cautious optimism. The virtue ethicist should therefore resist the idea that damage limitation exhausts our capacity for control. The alternative model of cognition suggests that indirect mitigation strategies aimed at short-term behavioural control may gradually reduce the underlying implicit biases themselves. Moreover, the most important items in our egalitarian toolboxes should be positive strategies aimed directly at enabling us to inhibit the cognitive effects of our biased associations rather than merely their behavioural influence. The ultimate aim should be weakening or eradicating implicit bias from response-directed processing.

Before turning to the defence of virtue ethics, I need to introduce the alternative model of cognition on which it depends. Section 3 explains this model in the context of psychological research on attitudes, and Section 4 outlines work on goal automaticity.

3 Attitudes in a Cognitive-Affective Personality System

The case for cautious optimism appeals to social psychologists' work on attitudes; dynamic, associative models of personality; and cognitive-affective processing. Although I cannot do justice to the literature here, this section highlights the most relevant aspects of the overall picture which emerges for the purpose of this chapter. In particular, the associative model of cognitive-affective personality can accommodate data cited in support of dual-process theory without the need to postulate separate systems for automatic and deliberative processing. Unlike models which posit distinct systems, therefore, this alternative can straightforwardly accommodate evidence that deliberation influences automatic cognition.

Maio and Haddock (2010) explain the social psychologist's conception of attitudes as complex, structured evaluations of objects with cognitive, affective, and behavioural components which have functional roles in a person's psychology. Attitudes are associative clusters of mental items which differ in content and strength. Their objects may be as particular and concrete as a drip of candle wax or as general and abstract as universal justice.

The *content* of one's attitude towards an object is a function of cognitive elements one associates with it such as a belief that woollen jumpers are difficult to wash; associated affective elements such as a fear of sheep; and associations with past behaviours such as the memory that one preferred wool to acrylic last time one bought a jumper. This last, behavioural factor is not so much a 'component' of the attitude as philosophers might understand it, but rather a trigger for attitude formation. In the absence of an existing, accessible attitude towards woollen jumpers, I may infer a positive attitude from my awareness of my past purchasing decisions. As work on cognitive dissonance shows, I may also alter an existing attitude as a result of attitude-incongruent behaviour (Cooper, 2007). For example, my purchase of one might lead me to adjust a negative attitude towards woollen jumpers. It is important that this reduction in negativity is a change in the *content* of the attitude and not, as philosophers might be inclined to say, in its strength. How much I like or dislike an attitude object is part of the content of that attitude.

In the context of a dynamic, associative model of personality, an attitude's *strength* is a matter of the strength of the connections between its components and with other elements in the cognitive-affective personality system (CAPS), situational features, behaviours, and so on. The strength of the connection between two components is a matter of how readily each affects the other's influence on cognition. As Mischel and Shoda (1995) explain it, the personality system involves five general types of 'cognitive-affective unit'. First, people classify features of internal and external experience using categories such as 'philosophy' and 'penguin'. Second, individuals have beliefs about themselves and their worlds such as 'I am going to mess up this job interview' and 'oil-soaked penguins need woollen jumpers'. Third, individuals' experience is affectively laden with such things as sympathy and claustrophobia. Fourth, people value aspects of their worlds such as patience and penguins. Fifth, individuals have plans and strategies such as intentions to assuage feelings of disappointment by thinking positively and to respond to the next oil spill by knitting penguin-sized woollen jumpers.

The various cognitive-affective units in the CAPS model are part of a connectionist network which processes cognitive and affective information and which is itself modified by that processing. Internal and external inputs such as the memory of rescue workers appealing for penguin-sized woollen jumpers or the sudden discovery of an ambiguous figure slumped on the corner of Miskin Street affect the flow of information across the network in two ways. First, they induce processing aimed at an immediate response such as intending to purchase wool or dialling 999. Second, this processing strengthens network connections between activated components.[1]

[1] The best way to understand this system is to study Mischel and Shoda's diagram (1995: 253).

Stronger attitudes are more *accessible* in the sense that they are more likely to significantly affect cognitive-affective processing, intention formation, and behaviour. Attitudes are strengthened and made more accessible by activation. The more often an attitude influences cognition, the stronger the associations between its components and the stronger the connections between those components and triggering internal and external elements. Accessibility in this sense need not be conscious. Processing units can be triggered *automatically* by external and internal stimuli, feedback, and associations. Processing can take place consciously or non-consciously, with or without an agent's awareness. That is, the associative model of cognitive-affective personality can accommodate data cited in support of dual-process theory without postulating distinct systems for automatized and deliberative cognitive processing. This is important because it allows the model to straightforwardly accommodate evidence for the influence of consciously endorsed commitments and deliberation on automatized cognition. For example, the associative model can more easily explain why forming an implementation intention to associate women with science or Muslims with peace especially quickly reduces bias on even implicit measures (Webb, Sheeran, and Pepper, 2012).

4 Goals, Attitudes, and Automaticity

Work on goal automaticity provides further evidence for the influence of deliberative cognition on automatized processing. Goals and commitments which are initially chosen as the result of conscious deliberation may become *automatized* if repeatedly invoked in cognitive processing, or they may result from an entirely automatic process. As with attitudes, a goal's accessibility is a matter of how readily it influences cognition, and goals are strengthened and made more accessible by activation (Bargh and Williams, 2006: 2). Like attitudes, goals are understood as located in an associative cognitive system which encompasses automatized perceptual sensitivities, proactive as well as reactive goals and motives, affective processing and more (Bargh, 1989, 1990, 2006: 147–8; Bargh, Gollwitzer, et al., 2001: 1014; Isen and Diamond, 1989).

The ability of complex, abstract goals to guide cognition automatically demonstrates the potential intelligence of automaticity. For example, temporarily raising the accessibility of the goal of cooperation outside conscious awareness caused subjects to behave as cooperatively as those explicitly asked to cooperate and significantly more cooperatively than controls (Bargh, Gollwitzer, et al., 2001). Bargh's work has explored the automatic activation of, and behavioural guidance by, 'higher-order goals and motives' relevant to social interaction such as commitments to truth, justice, and 'being a good mother, a high achiever, or a moral person' (Bargh, 1990: 103–4, 118; Bargh and Gollwitzer, 1994: 79).

Goals guide by associating features of situations with flexible and intelligent response strategies. As they are repeatedly activated, these associations become automatized. If one consciously selects the goal of cooperation sufficiently often in response to tensions over a shared resource, one will gradually associate such tensions with this response. That is, one's goal will initiate cooperation in response to such tensions without the need for conscious deliberation. The response is flexible, since it must be sensitive to the details of particular cases, avoiding not only trampling others' interests, but blocking others' attempts to trample one's own. The response is intelligent since it embodies one's reflective judgement about the best way to navigate a tricky aspect of one's social world.

Although both goals and attitudes can be automatized through habituation, they differ in their relation to acts of volition: goals, but not attitudes, are potential objects of deliberative choice. Although one cannot decide to automate a goal, one can decide to consciously adopt it, potentially beginning the process of automatization if conditions are right (Bargh and Williams, 2006: 2). This volitional distinction is of crucial importance to both individuals concerned about implicit bias and virtue ethicists. Since we cannot generally choose to like or dislike something by a mere act of will, even our explicit attitudes lie largely outside our direct control. If I am fond of penguins, I cannot just decide to dislike them, even though I could try to change my attitude indirectly by researching their less endearing habits. In particular, even our conscious associations are largely outside direct deliberative control. I cannot just decide to eliminate the association between penguins and winter festivities from my cognitive processing system. Since we have little direct control over even associations of which we are fully aware, there is likely to be little point in trying to eradicate implicit associations directly. Trying to 'will away' our implicit biases—or urging others to do so—is likely to be pointless at best and counterproductive at worst. In contrast, adopting a goal or making a commitment is precisely the sort of thing that acts of volition are good for.

In Section 5 I examine the effects of enduring, automatized egalitarian commitments on the expression of implicit bias and argue that the automatization of egalitarian goals has a key role to play in responses to implicit bias. While the familiarity of failed resolutions is indicative of the difficulties people experience in following through on their commitments, the effectiveness of implementation intentions suggests that psychological research could guide the selection of more successful strategies. Although they are too specific to fully capture the content of most commitments, implementation intentions might be incorporated into an effective overall strategy of goal pursuit. Further research on attitudes, goals, and specific cognitive strategies should enable us to better understand how to effectively habituate and maintain our commitments, enabling individuals to resist threats to goal pursuit and helping communities to encourage and sustain their commitments to egalitarianism.

5 Automatizing the Egalitarian's Toolbox

Since automatization systematically tunes the cognitive-affective processing system to reflect deliberatively endorsed values and commitments rather than working around its deficiencies, concerned individuals and virtue ethicists have good reason to be interested in the automatization of egalitarian commitments. The process of automatizing the goal of treating people fairly, for example, is precisely aimed at sensitizing the system to the right reasons and desensitizing it to the wrong ones. This is just the kind of habituation required for the development of virtue. In this section I focus on the habituation of egalitarian virtue. In Section 7 I explain why Mischel and Shoda's associative model of personality suggests that the habituation of egalitarian virtue should also dehabituate anti-egalitarian vice by decreasing implicit biases themselves.

I argue that the virtue ethicist can respond to the challenge of implicit bias by counselling the implicitly biased to habituate egalitarian virtues by adopting and pursuing egalitarian commitments. I develop this response by introducing two research programmes concerned with the effectiveness of such commitments. This research supports two claims: first, consciously chosen egalitarian commitments can be automatized; second, habituated egalitarian motivations can effectively guide automatic cognition.

Just as implicit and explicit bias are distinguished by the measures used to assess them (see the Introduction in Volume I), so it is with implicit and explicit egalitarian commitments. The first research programme I discuss concerns the differential effectiveness of different explicit motivations to avoid prejudice. The second concerns the effectiveness of implicit egalitarian commitments. In Section 6 I explore the differences between the psychological constructs posited by each programme, and explain why attempting to understand the automatization of egalitarian commitments in the light of both raises some puzzling questions.

I begin by outlining the different effects of two distinct kinds of explicit egalitarian commitment on expressions of prejudice. Individuals who are personally committed to not being prejudiced, as opposed to wishing to avoid appearing prejudiced, effectively avoid expressing prejudice even in ways which elude conscious control. I then outline evidence that the ability of such individuals to inhibit the influence of implicit stereotypes on cognition depends on automatization. This ability is especially significant because implicit stereotyping is more resistant to amelioration than other forms of implicit bias (Amodio, Devine, and Harmon-Jones 2008: 63).

Plant and Devine's (1998) 'Internal and external motivation to respond without prejudice' scales assess individual differences in kind and degree of egalitarian motivation. 'External' motivation stems from a concern with self-presentation: the individual wishes to avoid others' disapproval of prejudiced behaviour (EMS). 'Internal' motivation stems from a concern about prejudice itself: the individual's values are inconsistent with prejudice, and not being prejudiced is considered personally important (IMS). High-IMS individuals are motivated to avoid prejudice even when unobserved, and their egalitarian commitments are relatively consistent across different situations (Amodio, Devine, and Harmon-Jones 2008: 61). In contrast, low-IMS high-EMS individuals are motivated to respond without prejudice only in public scenarios, while low-IMS low-EMS individuals are not concerned to avoid expressions of prejudice at all.

What about differences *among* high-IMS individuals? One might think that high-IMS high-EMS individuals would demonstrate the least bias of all groups, since they have not one, but two, sources of motivation. In fact, however, high-IMS low-EMS individuals show the least bias. Although relative to low-IMS individuals, all high-IMS individuals show similarly reduced bias in responses subject to deliberative control, only those low-EMS high-IMS demonstrate less bias on relatively uncontrollable implicit measures (Devine et al., 2002; Amodio, Harmon-Jones, and Devine, 2003).

Why should additional egalitarian motivation undermine individuals' efforts to control prejudice? Devine et al. (2002: 846) suggest two possible explanations. First, high-IMS low-EMS individuals might never have acquired implicit bias, whereas high-IMS high-EMS individuals might be trying to overcome biased response patterns. Second, high-IMS high-EMS individuals might be at an earlier stage in a process of overcoming bias than high-IMS low-EMS individuals. Models of internalization hypothesize external motivations as a necessary first step on the path to automatization. On this account, high-IMS low-EMS individuals are low-EMS because they no longer need the support of external motivations having more fully integrated egalitarian values into their sense of themselves. The process of internalization is one of habituating patterns of responsiveness and, in the case of egalitarian commitments, of breaking others (Amodio, Harmon-Jones, and Devine, 2003: 751). All high-IMS individuals are committed to this process, but those at different points in the process have different motivational mixes.

Subsequent research supports the second, developmental model. This is important, because it suggests that the adoption of explicit egalitarian commitments enables individuals to change their implicit motivations by automatizing control of implicit bias. Amodio, Devine, and Harmon-Jones (2008) have shown that high-IMS low-EMS, but not high-IMS high-EMS, individuals are able to inhibit the influence of implicit stereotypes on cognition through automatized conflict-monitoring. High-IMS low-EMS individuals have highly accessible, automatized egalitarian commitments which conflict with implicit stereotypes at an early enough stage of cognitive processing for the conflict-monitoring mechanism to be effective in signalling the need for increased response regulation automatically and non-consciously. In contrast, high-IMS high-EMS individuals have less accessible, less automatized egalitarian commitments which are more reliant on the deliberative control effective only later in cognitive processing. The conflict-monitoring mechanism is therefore unable to signal the need for increased response regulation because little cognitive conflict occurs at the earlier stage of processing.

Further support for the effectiveness of automatized egalitarian commitments is provided by Moskowitz et al., who showed that highly accessible, enduring egalitarian goals can inhibit the activation of stereotypes preconsciously (Moskowitz et al., 1999; Amodio, Devine, and Harmon-Jones, 2008: 71–2). Individuals with similarly non-prejudiced attitudes and equally accessible cultural stereotypes but stronger commitments to fairness inhibited the influence of stereotypes on cognition preconsciously.

The effectiveness of automatized egalitarian commitments not only supports a stronger defence of virtue ethics by showing that habituated egalitarian motivations can reliably guide cognitive processing without the need for ongoing deliberative control. Once automatized, egalitarian commitments also have significant practical advantages over strategies requiring conscious control. In addition to inhibiting the influence of implicit bias on more automated cognitive processing, automatized egalitarian commitments are relatively efficient in terms of cognitive resources, relatively unimpeded by the erosion of cognitive capacity which results from effortful deliberation, and therefore relatively immune to the potential for rebound which characterizes conscious efforts to suppress the effects of stereotypes on deliberation (Park, Glaser, and Knowles, 2008; Glaser and Knowles, 2008; Moskowitz et al., 1999: 181). For instance, cognitive depletion increases 'shooter bias' for individuals low, but not high, in implicit motivation to control prejudice (IMCP) (Park, Glaser, and Knowles, 2008). Furthermore, Park, Glaser, and Knowles argue that the character of this particular psychological construct makes their demonstration of the effectiveness of implicit egalitarian motivations especially reliable (416). IMCP is a measure of the strength of two implicit associations: first, that between prejudice and bad; second, that between self and prejudice. Individuals who are strongly motivated by concerns about self-presentation and who have highly effective generic regulative capacities would be expected to demonstrate strong associations between prejudice and bad whether they actually had such associations or not. These individuals would also be expected to demonstrate weak associations between self and prejudice for just the same reasons, however, and so would not be assessed as high IMCP. Only individuals relatively unconcerned about self-presentation or with relatively weak generic regulative capacities would be expected to demonstrate both of the strong associations required for high IMCP. This makes it likely that the relation between high IMCP and the ability to inhibit the influence of implicit bias on cognition is a specific effect of strong implicit egalitarian motivations. This does not mean that the effectiveness of automatized egalitarian commitments requires an implicit (or explicit) belief that one is prejudiced. As Glaser and Knowles point out, the finding that strongly associating prejudice with bad is enough to inhibit the influence of implicit bias, but that strongly associating self with prejudice is not, is just what one would expect. One can be motivated by an egalitarian goal (unprejudiced behaviour) whether one thinks one is currently prejudiced or not, but merely believing that one is prejudiced will fail to motivate behavioural regulation unless one disvalues prejudice (2008: 170).

Taken together, these two research programmes are good news for the virtue ethicist. A satisfactory response to the challenge posed by implicit bias depends on two things. First, it must be possible to embed reflectively endorsed egalitarian motivations in the cognitive-affective processing system through habituation. This is supported by evidence for the developmental model of egalitarian motivation from IMS/EMS research. Second, habituated egalitarian motivations must be able to effectively guide cognition outside conscious awareness. This is supported by evidence from work on the effectiveness of implicit egalitarian commitments. Moreover, the gradual internalization and automatization of conscious egalitarian commitments, and the effectiveness of highly accessible and automatized egalitarian goals, is just what Mischel and Shoda's model of personality and work on goal automaticity predicts.

6 Puzzles about Automatization

Despite the promise of automatized egalitarian commitments, however, the picture which emerges from current research raises some puzzling questions.

Plant and Devine's measures of IMS and EMS differ significantly from Glaser and Knowles's measure of IMCP. Whereas the former depend on self-report, the latter are assessed using implicit measures of association. Moreover, it is currently unclear how these constructs are related. Glaser and Knowles (2008: 170–1) found neither high-IMS alone nor high-IMS low-EMS to affect the relation between strength of race-weapons stereotype and shooter bias. Similarly, Park, Glaser, and Knowles (2008: 414) found IMS and EMS to be correlated with neither race-weapons stereotype nor shooter bias, and no relation between IMCP and IMS, EMS or IMS-EMS interaction.

These results are puzzling in several respects. If high-IMS low-EMS individuals inhibit the influence of bias on cognition via automatized conflict-monitoring as the IMS/EMS research suggests, why do they not inhibit the effects of race-weapons stereotypes on their responses in the shooter task? This discrepancy cannot be explained by differences in the kinds of implicit bias studied, because the discrepancy appears to affect studies specifically focused on stereotypes. Whereas IMS/EMS researchers have found high-IMS low-EMS individuals to inhibit stereotypes preconsciously (Moskowitz et al., 1999; Amodio, Devine, and Harmon-Jones 2008), IMCP researchers have found no relation between high-IMS low-EMS and stereotype inhibition (Glaser and Knowles, 2008; Park, Glaser, and Knowles, 2008). Why should high-IMS low-EMS inhibit stereotypes preconsciously in one research programme but not the other? Moreover, if high-IMS low-EMS individuals avoid biased responses even on measures which elude conscious control because they have largely succeeded in automatizing their egalitarian commitments, why do they not demonstrate a highly accessible negative association with prejudice? That is, although high-IMS low-EMS does not seem obviously predictive of a strong association between self and prejudice, it seems odd that it does not correlate with a strong association between prejudice and bad. Furthermore, the associative model of personality seems *prima facie* to rule out explaining the various results in terms of two distinct routes to bias reduction, especially since both the conflict-monitoring element of the IMS/EMS research project and work on IMCP have examined the influence of race-weapons stereotypes (cf. Glaser and Knowles, 2008: 171).

As Devine et al. suggest, longitudinal studies of the development of IMS and EMS are needed to establish how high-IMS low-EMS individuals avoid bias and what might assist high-IMS high-EMS individuals to effectively pursue their egalitarian goals, as well as features of the social environment which might encourage low-IMS individuals to identify with egalitarian values (Devine et al., 2002: 846). Given the apparent discrepancies between results from work on IMS/EMS and IMCP, however, further research to clarify the relationship between the various constructs developed in the psychological literature on egalitarian motivation will be equally important. Particularly useful might be work comparing factors which support and sustain individual development of high-IMS, low-EMS, and high-IMCP.

One possibility is that high-IMS low-EMS individuals have strong egalitarian *goals*, whereas high-IMCP individuals have strong egalitarian *attitudes* in the sense explained in Section 4. Although high-IMCP is described in terms of goals in the literature, these are indirectly inferred from measures of implicit associations between prejudice and bad, and between self and prejudice.

A second possibility is that the motivation for adopting egalitarian goals is what matters: there is a difference between being motivated to pursue a personally important goal and being personally motivated to pursue an important goal. Completing this chapter might be personally important to me without my disapproving of those with no interest in pursuing academic philosophy. In contrast, kindness might be an important moral value such that I am concerned not only to be kind myself but to encourage and approve kindness in others. Perhaps high-IMCP individuals show less shooter bias than high-IMS low-EMS individuals because they are committed to egalitarianism for different reasons: whereas the former are personally motivated to avoid prejudice because they disvalue it generally, the latter may see it as a merely personal project, albeit one they happen to care strongly about. Perhaps this explains why high-IMS low-EMS does not predict a strong negative association with prejudice as such.

A third possibility is that high-IMS low-EMS alone is insufficient for the formation of an enduring, highly accessible egalitarian goal. Amodio, Devine, and Harmon-Jones (2008: 63) included an experimental manipulation to increase the accessibility of subjects' internal, but not external, motivations to avoid prejudice. Although the automatized egalitarian values bound up with the self-concepts of high-IMS low-EMS individuals enabled them to control their responses, it is possible that the automatization of egalitarian goals is a further step. Individuals without such automatized goals might be insufficiently sensitive to situations requiring control, and so fail as a group to demonstrate less shooter bias when the need for control is not made especially salient. While one might expect the content of the shooting task to make such a need salient, perhaps this is undermined by the artificial form of the simulation.

7 Evaluating the Egalitarian's Toolbox

The defence of virtue ethics outlined so far appeals to the potential of automatized egalitarian commitments to inhibit the cognitive influence of implicit bias outside conscious awareness and without the need for deliberative control. A further advantage of habituated egalitarian commitments is their potential to eliminate or reduce our implicit biases themselves. The dynamic, associative model of personality developed by Mischel and Shoda suggests that enduring commitments to egalitarianism should decrease implicit bias itself as a long-term effect of the automatization which enables such commitments to prevent its influence on cognition.

Current research provides qualified support for this possibility. For example, Rudman, Ashmore, and Gary (2001) found that students who voluntarily enrolled in diversity education demonstrated reductions in implicit, as well as explicit, bias. The class was designed to increase students' awareness of racial prejudice and motivation to overcome racism in themselves, as well as providing the opportunity to make social contact with 'out-group' members in a safe and supportive atmosphere. The results suggested distinct, but mutually supportive, cognitive and affective routes to reductions in explicit and implicit bias respectively. Egalitarian commitments were plausibly partially responsible for these effects: a decision to enroll in diversity education suggests motivation to overcome prejudice (Rudman, Ashmore, and Gary, 2001: 866), and the class was designed to consolidate and support pursuit of this initial commitment. It is unfortunate that long-term results were not evaluated, since it would be useful to know if the two routes to bias reduction would converge over time, as Mischel and Shoda's model of personality predicts. Although cognitive and affective changes might differentially support reductions in explicit and implicit bias in the short-term, this model of cognition understands the two routes as affecting a single connectionist network. Given the short-term nature of the study, however, it is unsurprising that the two routes were only weakly correlated: the model predicts that change will be gradual and that adjustments to one element of the system (e.g. a belief) will affect associated elements of the system (e.g. an affective response) only slowly as the initial change repeatedly affects the flow of information across the system. Although such research suggests that interventions might plausibly 'kick-start' the process of automatization, therefore, further research is required to substantiate this possibility.

Moreover, this picture does not yet capture the complexity of psychological reality. Research suggests that the effect of automatized egalitarian commitments depends on the *kind* of implicit bias in question. Whereas egalitarian commitments seem to reduce or eliminate implicit affective bias, for example, they seem to inhibit rather than weaken implicit stereotypes (Amodio, Harmon-Jones, and Devine, 2003; Amodio, Devine, and Harmon-Jones, 2008: 63; Moskowitz et al., 1999).

Why should the kind of implicit bias matter if the cognitive-affective personality system is a connectionist network? One difference between affective bias and stereotypes is that an understanding of the social world seems to depend on the latter but not the former. Perhaps some features of the cognitive-affective system (e.g. implicit stereotypes) may be strongly connected to elements implicated in processing motivated by a need to understand, despite being weakly connected to elements implicated in processing motivated by a need to respond. So long as her society harbours stereotypes, a member of that society will need to be aware of them in order to effectively navigate her social world. The process of automatizing egalitarian goals might therefore weaken connections between stereotypes and elements of the cognitive system implicated in processing aimed at responding without weakening the stereotypes' connections with elements implicated in processing aimed at social understanding. Indeed, automatized egalitarian goals might be partially constituted by the preconscious inhibition of stereotypes in response-directed processing. If this were the whole story, the stereotypes themselves might be expected to weaken over time. Since stereotypes are crucial to understanding the social world, however, processing aimed at social understanding would continue to activate and sustain them. Because response-directed processing flows across the same network as processing aimed at understanding, the automatization of egalitarian goals would therefore tend to isolate stereotypes from response-directed processing without eliminating them.

8 Beyond Individual Commitment

I have argued that individual egalitarian commitments can play an essential part in resisting implicit bias and the challenge it presents to virtue ethics. But individuals' ability to sustain and implement their commitments depends crucially on hospitable environments. The process of automatization is designed to select and refine *successful* strategies and response patterns. We automatically adjust our strategies in response to their success or failure in enabling us to navigate social interactions. Although we can deliberatively adopt egalitarian goals independently of others, therefore, we have only limited control over our ability to automatize them, because the success of strategies aimed at achieving those goals depends on others' cooperation. While our ability to automatize intelligent and flexible responses can support individuals' pursuit of egalitarian commitments, therefore, this intelligence and flexibility also renders those commitments vulnerable to inhospitable social environments. In hospitable environments, egalitarian commitments will be strengthened by the process of modification and refinement which is essential to automatizing them, enabling individuals to habituate appropriately sensitive responses to pertinent features of their social environments. In hostile environments, however, this same process will tend to weaken and undermine individuals' egalitarian commitments, because the cognitive system will automatically modify and refine them in response to difficulties in implementing them, others' hostile responses, and unsuccessful social interactions. Community support for egalitarianism is, therefore, essential for the development and maintenance of egalitarian virtue not because there is nothing which would count as virtue in a prejudiced social environment, but because the human cognitive processing system is designed to automatically adapt our responses to whatever social environment we happen to inhabit. Communities and institutions which themselves embody egalitarian values, and which encourage more individuals to adopt egalitarian goals, are therefore crucial. Environmental and cultural interventions which foster and support strong commitments on the part of individuals, and which seek to make such commitments institutional and social norms, are thus essential to egalitarian toolboxes.

This echoes Aristotle's emphasis on the development and practice of individual virtue in the context of a supportive moral community (see Aristotle, 2002). The process of automatizing egalitarian goals outlined here just is the habituation of appropriate values and commitments. The internalization and automatization of egalitarian motivation is a process of tuning the cognitive-affective personality system to respond appropriately to just the right features of individuals' external social and internal psychological environments. That is, the habituation of appropriate responsiveness to egalitarian reasons is part of the development and practice of practical wisdom.

Acknowledgements

I am grateful to all those involved in the Implicit Bias and Philosophy International Research Project for supporting and stimulating my thinking on this topic, especially participants at the July 2012 workshop who provided feedback on the ideas which formed the basis for this chapter. I am particularly grateful to Jennifer Saul for supporting my participation in the project. I would like to thank Gregory R. Maio and Geoffrey Haddock for allowing me to attend their lectures on the social psychology of attitudes and for their patience in answering a philosopher's questions, and members of Cardiff's moral psychology reading group for ongoing discussions of the implications of empirical psychology for ethical theory. I owe special thanks to Michael Brownstein, Jennifer Saul, and Jonathan Webber for extensive and invaluable feedback on earlier drafts, and to the anonymous referees whose insights helped me to see where further clarification was needed.

References

Amodio, David M., Eddie Harmon-Jones, and Patricia G. Devine (2003). 'Individual differences in the activation and control of affective race bias as assessed by startle eyeblink response and self-report'. *Journal of Personality and Social Psychology* 84(4): 738–53.

Amodio, David M., Patricia G. Devine, and Eddie Harmon-Jones (2008). 'Individual differences in the regulation of untergroup bias: The role of conflict monitoring and neural signals for control'. *Journal of Personality and Social Psychology* 94(1): 60–74.

Annas, Julia (2011). *Intelligent Virtue*. Oxford and New York: Oxford University Press.

Aristotle (2002). *Nicomachean Ethics*. Translated, with a historical introduction, by Christopher Rowe. Philosophical introduction and commentary by Sarah Broadie. Oxford and New York: Oxford University Press.

Baier, Annette C. (1995). 'What do women want in a moral theory?'*Moral Prejudices: Essays on Ethics*. Cambridge, MA: Harvard University Press: 1–17. Revision of 'What do women Want in a moral theory?' *Noûs* 19(1) (1985): 53–63.

Bargh, John A. (1989). 'Conditional automaticity: Varieties of automatic influence in social perception and cognition'. In *Unintended Thought*, ed. James S. Uleman and John A. Bargh. New York and London: Guilford Press: 3–51.

Bargh, John A. (1990). 'Auto-motives: Preconscious determinants of social interaction'. In *Handbook of Motivation and Cognition: Foundations of Social Behavior*, vol. 2., ed. Richard M. Sorrentino and E. Tory Higgins. New York and London: Guilford Press: 93–130.

Bargh, John A. (2006). 'What have we been priming all these years? On the development, mechanisms, and ecology of nonconscious social behavior'. *European Journal of Social Psychology* 36: 147–68.

Bargh, John A. and Peter M. Gollwitzer (1994). 'Environmental control of goal-directed action: Automatic and strategic contingencies between situations and behavior'. In *Integrative Views of Motivation, Cognition and Emotion*, vol. 41, ed. William D. Spaulding. Nebraska Symposium on Motivation. Lincoln, NE, and London: University of Nebraska Press: 71–124.

Bargh, John A., Peter M. Gollwitzer, et al. (2001). 'The automated will: Nonconscious activation and pursuit of behavioral goals'. *Journal of Personality and Social Psychology* 81(6): 1014–27.

Bargh, John A. and Erin L. Williams (2006). 'The automaticity of social life'. *Current Directions in Psychological Science* 15(1): 1–4.

Baumeister, Roy F. et al. (1998). 'Ego depletion: Is the active self a limited resource?' *Journal of Personality and Social Psychology* 74(5): 1252–65.

Blair, Irene V., Jennifer Ma, and Alison Lenton (2001). 'Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery'. *Journal of Personality and Social Psychology* 81(5): 828–41.

Blank, Rebecca M. (1991). 'The effects of double-blind versus single-blind reviewing: Experimental evidence from the *American Economic Review*'. *American Economic Review* 81: 1041–67.

Brennan, Samantha (2009). 'Feminist ethics and everyday inequalities'. *Hypatia* 24(141): 141–59.

Budden, Amber E. et al. (2008). 'Double-blind review favours increased representation of female authors'. *Trends in Ecology and Evolution* 23(1): 4–6.

Casey, Pamela M. et al. (2012). *Helping Courts Address Implicit Bias: Strategies to Reduce the Influence of Implicit Bias*. National Center for State Courts. Summary of *Helping Courts Address Implicit Bias: Resources for Education*.

Cooper, Joel M. (2007). *Cognitive Dissonance: Fifty Years of a Classic Theory*. Los Angeles, CA: Sage.

Corcoran, Katja, Tanja Hundhammer, and Thomas Mussweiler (2009). 'A tool for thought! When comparative thinking reduces stereotyping effects'. *Journal of Experimental Social Psychology* 45(4): 1008–11.

Dasgupta, Nilanjana, and Shaki Asgari (2004). 'Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping'. *Journal of Experimental Social Psychology* 40(5): 642–58.

Dasgupta, Nilanjana and Anthony G. Greenwald (2001). 'On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals'. *Journal of Personality and Social Psychology* 81(5): 800–14.

Devine, Patricia G. et al. (2002). 'The regulation of explicit and implicit race bias: The role of motivations to respond without prejudice'. *Journal of Personality and Social Psychology* 82(5): 835–48.

Doris, John Michael (2009). 'Skepticism about persons'. *Philosophical Issues* 19: *Metaethics*: 57–91.

Equality Challenge Unit (2013). <https://blogs.shu.ac.uk/internationalnetwork/2013/10/30/equality-link-october/>.

Galinsky, Adam and Gordon B. Moskowitz (2000). 'Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism'. *Journal of Personality and Social Psychology* 78(4): 708–24.

Glaser, Jack and Eric D. Knowles (2008). 'Implicit motivation to control prejudice'. *Journal of Experimental Social Psychology* 44(1): 164–72.

Goldin, Claudia and Cecilia Rouse (2000). 'Orchestrating impartiality: The impact of "blind" auditions on female musicians'. *American Economic Review* 90(4): 715–41.

Isen, Alice M. and Gregory A. Diamond (1989). 'Affect and automaticity'. In *Unintended Thought*, ed. James S. Uleman and John A. Bargh. New York and London: Guilford Press: 124–52.

Jost, John T. et al. (2009). 'The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore'. *Research in Organizational Behavior* 29: 39–69.

Kamtekar, Rachana (2004). 'Situationism and virtue ethics on the content of our character'. *Ethics* 114(3): 458–91.

Kelly, Daniel, Edouard Machery, and Ron Mallon (2010). 'Race and racial cognition'. In *The Moral Psychology Handbook*, ed. John Michael Doris and The Moral Psychology Research Group. Oxford: Oxford University Press: 433–72.

Kelly, Daniel and Erica Roedder (2008). 'Racial cognition and the ethics of implicit bias'. *Philosophy Compass* 3(3): 522–40.

Lerner, Jennifer S. and Philip E. Tetlock (1999). 'Accounting for the effects of accountability'. *Psychological Bulletin* 125(2): 255.

Maio, Gregory R. and Geoffrey Haddock (2010). *The Psychology of Attitudes and Attitude Change*. Los Angeles, CA: Sage.

Merritt, Maria W. (2009). 'Aristotelean virtue and the interpersonal aspect of ethical character'. *Journal of Moral Philosophy* 6: 23–49.

Merritt, Maria W., John M. Doris, and Gilbert Harman (2010). 'Character'. In *The Moral Psychology Handbook*, ed. John M. Doris and The Moral Psychology Research Group. Oxford: Oxford University Press: 355–401.

Milgram, Stanley (2009). *Obedience to Authority: An Experimental View*. Intr. Philip G. Zimbardo. New York: HarperCollins.

Mischel, Walter, and Yuichi Shoda (1995). 'A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure'. *Psychological Review* 102(2): 246–68.

Moskowitz, Gordon B. et al. (1999). 'Preconscious control of stereotype activation through chronic egalitarian goals'. *Journal of Personality and Social Psychology* 77(1): 167–84.

Muraven, Mark and Roy F. Baumeister (2000). 'Self-regulation and depletion of limited resources: Does self-control resemble a muscle?' *Psychological Bulletin* 126(2): 247–59.

Muraven, Mark, Dianne M. Tice, and Roy F. Baumeister (1998). 'Self-control as limited resource: Regulatory depletion patterns'. *Journal of Personality and Social Psychology* 74(3): 774–89.

Park, Sang Hee, Jack Glaser, and Eric D. Knowles (2008). 'Implicit motivation to control prejudice moderates the effect of cognitive depletion on unintended discrimination'. *Social Cognition* 26(4): 401–19.

Pelham, Brett, Matthew Mirenberg, and John Jones (2002). 'Why Susie sells seashells by the seashore: Implicit egotism and major life decisions'. *Journal of Personality and Social Psychology* 82(4): 469–87.

Peters, Douglas P. and Stephen J. Ceci (1982). 'Peer-review practices of psychological journals: The fate of published articles, submitted again'. *Behavioral and Brain Sciences* 5(2): 187–95.

Plant, E. Ashby and Patricia G. Devine (1998). 'Internal and external motivation to respond without prejudice'. *Journal of Personality and Social Psychology* 75(3): 811–32.

Quinn, Andrew and Barry R. Schlenker (2002). 'Can accountability produce independence? Goals as determinants of the impact of accountability on conformity'. *Personality and Social Psychology Bulletin* 28(4): 472–83.

Rees, Clea F. and Jonathan Webber (2014). 'Automaticity in virtuous action'. In *The Philosophy and Psychology of Virtue: An Empirical Approach to Character and Happiness*, ed. Nancy E. Snow and Franco V. Trivigno. New York and London: Routledge: 75–90.

Remploy (2013). 'Changing perceptions with changing faces'. <http://www.remploy.co.uk/info/20137/partners_and_programmes/168/changing_faces>.

Rudman, Laurie A., Richard Ashmore, and Melvin Gary (2001). '"Unlearning" automatic biases: The malleability of implicit prejudice and stereotypes'. *Journal of Personality and Social Psychology* 81(5): 856–68.

Russell, Nestar J. C. (2011). 'Milgram's obedience to authority experiments: Origins and early evolution'. *British Journal of Social Psychology* 50(1): 140–62.

Saul, Jennifer (2012). 'Ranking exercises in philosophy and implicit bias'. *Journal of Social Philosophy* 43(3): 256–73.

Snow, Nancy E. (2009). *Virtue as Social Intelligence: An Empirically Grounded Theory*. New York: Routledge.

Webb, Thomas L., Paschal Sheeran, and John Pepper (2012). 'Gaining control over responses to Implicit Attitude Tests: Implementation intentions engender fast responses on attitude-incongruent trials'. *British Journal of Social Psychology* 51(1): 13–32.

3.2

Context and the Ethics of Implicit Bias

Michael Brownstein

1 Introduction

Early research on implicit attitudes and implicit biases emphasized the "directness" of the link between apparent triggers of those attitudes and behavior. For example, John Bargh and colleagues argued that there is a direct link between the perception of cues relevant to one's implicit attitudes and behavior. They write (1996: 231): "…social behavior is often triggered automatically on the mere presence of relevant situational features." The implication of this view is that merely being in the presence of a member of a socially stigmatized group may be sufficient to activate implicit attitudes and to cause one to act in biased ways.[1] In a related vein, Patricia Devine (1989) argued in a seminal paper that scores on measures of implicit bias reflect mere knowledge of cultural stereotypes. She showed that both egalitarians and non-egalitarians associate blacks with negative stereotypes, and suggested that there is a direct relationship between merely having cultural knowledge of stereotypes and behaving in biased ways.[2]

More recently, though, context has been shown to significantly affect the activation and expression in behavior of implicit attitudes. Some philosophers have begun to consider the ramifications of this development in the empirical literature. Understandably, most who have considered the effects of context on implicit attitudes have focused on the metaphysical ramifications of this fact; in particular, whether implicit attitudes represent a singular kind, or rather, whether the implicit attitude construct stands for a number of related (or perhaps

[1] See also Dijksterhuis and Bargh (2001).

[2] See Nosek and Hansen (2008) for discussion.

unrelated) cognitive and affective processes.[3] In this chapter I focus on a separate philosophical issue stemming from findings about the effects of context on implicit attitudes. There are, I will argue, ethical ramifications stemming from these findings. Any comprehensive "ethics of implicit bias," I will argue, must focus on outlining how agents can cultivate the right sort of relationships with the situations and contexts that affect their attitudes and behavior. This notion, of cultivating the right sort of "ambient" relationships, has been underdescribed by most ethical thinking about implicit bias, which usually focuses on the relationship between attitudes or mental states *within* agents.

First, I will outline recent data on the effects of context on the activation and expression in behavior of implicit biases (Section 2). I will then briefly describe the predominant ways of thinking about the ethics of implicit bias in the philosophical literature: the "ethics of internal harmony" (Section 3.1); the "world-first" strategy (Section 3.2); and the "seek/avoid" strategy (Section 3.3). Each of these conceptualizations reflects a crucial facet of the ethics of implicit bias, but each on its own is limited. I go on to describe a "contextualist" approach which incorporates elements of each of these by focusing on the relationship between agents' internal states and their ambient environment and situations (Section 4). The "nodes" of this relationship between agents and their environments are various forms of contextual cues, I argue. In focusing on these cues, agents can start on the road toward creating the kind of world that promotes ethical thought and action within themselves and others.

2 Context and Implicit Bias

In a recent paper, Bertram Gawronski and Joseph Cesario (2013) argue that implicit attitudes are subject to "renewal effects," which is a term used in animal learning literature to describe the recurrence of an original behavioral response after the learning of a new response. As Gawronski and Cesario explain, renewal effects usually occur in contexts other than the one in which the new response was learned. For example, a rat with a conditioned fear response to a sound may have learned to associate the sound with an electric shock in context A (e.g. its cage). Imagine then that the fear response is counterconditioned in context B (e.g. a different cage). An "ABA renewal" effect occurs if, upon being placed back in A (its original cage), the fear response returns. Gawronski and Cesario argue that implicit attitudes are subject to renewal effects like these.

[3] See, for instance, Machery (Volume I); Holroyd and Sweetman (Volume I); Madva and Brownstein (ms.); and discussion between Mazarin Banaji and Tamar Gendler on "The Mind Report." <http://bloggingheads.tv/videos/15811>.

For example, a person might learn biased associations while hanging out with their friends (context A), effectively learn counterstereotypical associations while taking part in a psychology experiment (context B), then exhibit behaviors consistent with biased associations when back with their friends. Gawronski and Cesario discuss several studies (in particular, Rydell and Gawronski, 2009, and Gawronski et al., 2010) that demonstrate in controlled lab settings renewal effects like these in implicit attitudes. The basic method of these studies involves an impression formation task in which participants are first presented with valenced information about a target individual who is pictured against a background of a particular color. This represents context A. Then, the same target individual is presented with oppositely valenced information against a background of a different color, representing context B. Participants' evaluations of the target are then assessed using an affective priming task in which the target is presented against the color of context A. In this ABA pattern, participants' evaluations reflect what they learned in context A. Gawronski and Cesario report similar renewal effects (i.e. evaluations consistent with the valence of the information that was presented first) in the patterns AAB (where the original information and new information are presented against the same background color A, and evaluations are measured against a novel background B), and ABC (where original information, new information, and evaluation all take place against different backgrounds).

These studies suggest that minor features of agents' context—like the background color against which an impression of a person is formed—can influence the activation of implicit attitudes, even after those attitudes have been "unlearned." These are striking results, but they are consistent with a broad array of findings about the influence of context and situation on the activation and expression in behavior of implicit attitudes. Context is, of course, a broad notion. Imagine taking an Implicit Association Test (IAT; Greenwald et al., 1998). The background color against which the images of the target subjects are presented is part of the context. Or perhaps before administering the IAT, the experimenter asks the subject to imagine herself as a manager at a large company deciding whom to hire. Having imagined oneself in this powerful social role will have effects on one's implicit evaluations (see the discussion later in this section), and these effects too can be thought of as part of one's context. Similarly, perhaps one had a fight with one's friend before entering the lab, and began the task feeling an acute sense of disrespect, or was hungry and jittery from having drunk too much caffeine, and so on . . .[4]

[4] Perhaps it would be better to refer to all of these various elements as "situational factors" rather than elements of context. See Section 3.3 for discussion.

As I will use the term, “context” can refer to any stimulus that moderates the way an agent evaluates or responds behaviorally to a separate conditioned stimulus. Anything that acts, in other words, as what theorists of animal learning call an “occasion setter” can count as an element of context.[5] A standard example of an occasion setter is an animal’s cage. A rat may demonstrate a conditioned fear response to a sound when in its cage, but not when in a novel environment. Similarly, a person may feel or express biased attitudes toward members of a social group only when in a particular physical setting, when playing a certain social role, or when feeling a certain way. To briefly review the extant data, I will note three kinds of contextual elements—perceptual, conceptual, and motivational—that have been shown to influence the activation and expression in behavior of implicit attitudes. These three kinds of contextual elements overlap significantly, of course. I carve them up this way only as a device for presentation of the empirical literature.

Perceptual elements of context influence implicit attitudes primarily in light of their visual, aural, or other sensory properties. For example, Mark Schaller and colleagues (2003) found that the relative darkness or lightness of the room in which participants sit shifts scores of implicit racial evaluations across several indirect measures, including the IAT.[6] The sort of renewal effects which Gawronski and Cesario discuss (above) are also examples of perceptual elements of context influencing the activation of implicit attitudes.

Conceptual elements of context influence implicit attitudes primarily in light of perceived group membership and social roles. Jamie Barden and colleagues (2004), for example, varied the category membership of targets by presenting the same individual in a prison context dressed as a prisoner and dressed as a lawyer; implicit evaluations of the person dressed as prisoner were considerably more negative. Similarly, Jason Mitchell and colleagues (2003) showed that implicit evaluations of the same individual—Michael Jordan—depended on whether he was categorized by race or occupation. Conceptual elements of context also include one’s own social role. Ana Guinote and colleagues (2010), for example, led participants in a high-power condition to believe that their opinions would impact the decisions of their school’s “Executive Committee,” and these participants showed more racial bias on both an IAT and an Affect Misattribution Procedure (AMP; Payne et al., 2005) than those in a low-power condition,

[5] I am indebted to Gawronski and Cesario (2013) for the idea that context acts as an occasion setter. On occasion setting and animal learning, see Schmajuk and Holland (1998).

[6] These results obtained only for subjects with chronic beliefs in a dangerous world.

who were led to believe that their opinions would not affect the committee's decisions.[7]

Finally, motivational elements of context influence implicit attitudes primarily in light of fluctuations in mood and emotion. Nilanjana Dasgupta and colleagues (2009), for instance, found that salient emotions selectively influence the activation of implicit attitudes. Participants who were induced to feel disgust had more negative evaluations of homosexuals on an IAT, although their implicit evaluations of Arabs remained unchanged. However, participants who were induced to feel anger had more negative evaluations of Arabs, while their evaluations of homosexuals remained unchanged.[8] In a related vein, Jennifer Kubota and Tiffany Ito (2014) recently showed that the emotional expression of others' faces (e.g. smiling) moderates the activation of black-danger stereotypes.

These data should not be surprising. For example, it is not hard to imagine a person who treats her colleagues fairly regardless of race, but (unwittingly) grades her students unfairly on the basis of race. Perhaps being in the superordinate position of professor activates this person's prejudices while being in an equal-status position with her colleagues does not. Perhaps perceptual and motivational factors play a role too. There could be posters on the wall of her classroom for films that propagate stereotypes, and these posters might render her stereotypical associations more psychologically available. As might simply being in a bad mood, or not getting enough sleep.

3 Three Ethics of Implicit Bias

What follows are three ways of thinking about the ethics of implicit bias. Each is connected to the others and is independently plausible. However, each on its own

[7] In the case of implicit gender bias, Richeson and Ambady (2001) found that assigning male participants to a subordinate role in a dyadic interaction with a woman led the men to have more negative implicit evaluations of women on an IAT, while male participants assigned to an equal-status or superordinate role showed favorable evaluations of the women. This result is interesting to compare to Guinote and colleagues' findings. Both studies find effects of one's perceived status on one's implicit intergroup attitudes; but in the case of race, perceiving oneself in a superordinate status seems to amplify bias, while in the case of gender, perceiving oneself in a subordinate status seems to amplify bias.

[8] See Gawronski and Sritharan (2010) for summary and discussion of these data. For more on context and implicit bias, one could look to the literature on the relationship between context and habit (on the plausible assumption that the behavioral upshots of implicit bias are habit-like). Neal and colleagues (2011) showed, for example, that people who habitually eat popcorn at the movies will be minimally influenced by hunger or by how much they like what they are eating, but only when the right physical context-cues obtain. People eat less out of habit if they are not in the right place—e.g. a meeting room rather than a cinema—or if they cannot eat in their habitual way—e.g. if forced to eat with their non-dominant hand.

faces significant theoretical and practical challenges. The way forward includes elements of each and connects them by focusing on the concept of context, to which I return in Section 4.

3.1 The ethics of internal harmony

One reason implicit biases are ethically pernicious is that in many cases they tend to persist and to influence the behavior of individuals who do not endorse the validity of those very biases (Gawronski and Bodenhausen, 2006; Nosek and Hansen, 2008). They can be aversive with respect to agents' reflective or moral commitments; in other words, giving rise to "aversive racism" (Dovidio and Gaertner, 2000, 2004). Philosophers writing on implicit bias have focused on this fact, and for good reason.[9] The idea that implicit biases can persist and influence the behavior of individuals who disavow them raises important ethical questions (not to mention metaphysical and epistemological questions). For example, Tamar Szabó Gendler (2008b) describes the ethical conflict arising from the pervasiveness of implicit bias in terms of agents being put into a state of "internal disharmony." This state is the result of discord between one's "aliefs"—which for all relevant purposes here we can think of as one's implicit attitudes—and one's beliefs. The relevant ideal to which one can aspire in order to combat internal disharmony is, of course, internal harmony. This is an ideal with a long history, stretching all the way back to Plato, who claimed that a just person is one who "puts himself in order, harmonizes . . . himself . . . [and] becomes entirely one, moderate and harmonious . . . " (*Republic*: 443de in Plato, 380 BCE/1992; quoted in Gendler, 2008b: 572).

This ideal of internal harmony—or something quite like it—is also found in the psychological literature. Keith Payne and Daryl Cameron (2010: 445), for example, write: " . . . the message of implicit social cognition is that the thoughts people introspect and report about do not tell the whole story of why they believe the things they believe and why they do the things they do." It is important that people learn this message, furthermore, because implicit cognition, like life in Hobbes' state of nature, can be "nasty, brutish, and short-sighted" (2010: 445). When we learn the facts about implicit social cognition—in particular, how it "can cause our ethicality to corrode"—we can "engage [in] better moral self-regulation in pursuit of our ideals" (2010: 456). Payne and Cameron's chapter in the *Handbook of Implicit Social Cognition* even begins with this epigraph from

[9] See, for instance, Gendler (2008a, 2008b); Madva (2012); Huebner (2009); and Kelly and Roedder (2008).

Rousseau: "Virtue is a state of war, and to live in it we have always to combat with ourselves."

The "ethics of internal harmony" addresses at least one element of what is undeniably frightening about implicit bias: namely, that it can affect one's own attitudes and behavior, even if one genuinely desires to be unbiased. However, the ethics of internal harmony faces challenges in representing an ethical response to the problem of implicit bias. First, the ideal of internal harmony is too permissive. A person with implicit biases, who is *also* explicitly biased, will (formally, at least) count as internally harmonious. Her ideals—as represented by her explicit attitudes—will line up with her implicit attitudes, as well as with those behaviors affected by her implicit attitudes. Minimally, this problem represents a *prima facie* challenge to the ethics of internal harmony.

More importantly, the ethics of internal harmony is predominantly *agent-centered*. It is agent-centered in the sense that it singularly recommends self-regulatory effort aimed at controlling or changing mental and emotional states internal to agents. Internal harmony describes an occurrent or dispositional status of one's own cognitive and emotional states. Gendler considers two strategies for regulating one's "belief-discordant aliefs" in the hopes of becoming more internally harmonious (2008b: 554): the "cultivation of alternative habits through deliberate rehearsal" and "refocusing of attention through directed imagination." While extremely valuable, both of these strategies focus exclusively on the agent, and in two senses. First, both the cultivation of habits and the refocusing of attention are things one does to oneself; the object of attention and regulatory effort are one's own habits or one's own patterns of attention. Second, both the cultivation of habits and the refocusing of attention are things one does in order to create harmony between one's implicit and explicit attitudes, both of which "belong to" the agent.

Others have articulated problems with an agent-centered ethics like this, problems having to do with the relative importance of social, institutional, and economic inequality (e.g. Huebner, 2009, Volume I: Anderson, 2010; Haslanger, 2015; Jacobson, this volume). I discuss this "institutional" critique of the self-regulation of implicit bias in Section 3.2. A related problem with an agent-centered ethics, not explicitly discussed in these critiques, is that it can have a framing effect on thought and action. For example, the overarching moral problem of implicit bias is not that it causes well-intentioned egalitarians to be internally disharmonious. The overarching moral problem is that implicit biases perpetuate injustice. The ethics of internal harmony is not formally inconsistent with this point, of course. One can aim to fight against injustice precisely by bringing one's implicit attitudes and behavior in line with one's reflective ends. But as an ethical end unto itself, the ideal of internal harmony frames one's attention predominantly on oneself in such a way that it risks overshadowing the greater moral problem. For example, one might think one's ethical "work" is finished once one has reached a state of internal harmony. But in a world suffused with injustice, this would be an unsatisfying end-point for ethics. It does not address one's obligations to encourage one's peers to act in ethical ways, for instance. Nor does it commit anyone to changing the institutional structures that help to perpetuate injustice, such as non-anonymous review of student papers, job application materials, and so on.

3.2 The world-first strategy

Those who have stressed the importance of changing social, institutional, and economic inequalities in order to combat implicit bias (Huebner, 2009, Volume I; Anderson, 2010; Haslanger, 2015; Jacobson, this volume) have articulated what I call a "world-first" strategy. On this view, the task of an ethics of implicit bias is to change institutions, not individuals. Generally, the idea behind this argument is that prejudice and bias are sustained by material forms of inequality, and that so long as social institutions continue to promote these forms of inequality, no amount of self-regulation can successfully combat the pernicious effects of racism, sexism, and other prejudices on one's own behavior. Bryce Huebner (2009: 88), for example, argues that "the only way in which we will be able to adequately modify our psychology is by modifying the world in which we live."[10] Huebner's view is that in a propaganda-filled world like ours, where even the most harmonious agent continues to be bombarded with stereotypes, any ethical strategy focused on making change at the individual level amounts to too little, too late.

The world-first strategy is appealing for several reasons. First, it connects ethical and moral concerns, by focusing one's attention "outward," off oneself and onto the world in which injustice is perpetrated. Second, I think proponents of the world-first strategy are correct in worrying that implicit biases may be relearned in social environments suffused with inequality.[11] Third, the world-first strategy serves as an important call for activism. Implicit bias is, of course, just one element of wider patterns of discrimination. Fighting to change housing laws, pay discrimination, racial profiling, and so on, can

[10] Huebner does discuss some self-regulatory strategies for modifying one's psychology, but he is skeptical of their effectiveness without an attendant revolutionary politics. See also Dixon et al. (2012), who voice related worries about prejudice reduction and attitude change. For a defense of "psychological" approaches to combating implicit bias, see Machery et al. (2010).

[11] Although see Madva (ms.) for some important doubts about this "relearning" worry.

therefore serve the cause of combating implicit bias as well as other social–moral travesties.

There are weak and strong interpretations of the world-first strategy, however. On a weak interpretation, it suggests that one ought to *both* try to change one's own attitudes and try to change the world itself. On a strong interpretation, trying to change one's own attitudes through self-regulation is hopeless, because biases will inevitably be relearned in an unjust social world; or, trying to change one's own attitudes is problematically distracting from the "real" work of changing the social, institutional, and economic sources of bias and discrimination.

The weak interpretation of the world-first strategy is more appealing than the strong interpretations. One reason for this is that there are simply not yet enough data to confirm or deny the charge of hopelessness. Research on the self-regulation of implicit bias is in its relative infancy. There is reason at least to think that this research is promising.[12] Second, it is hard to see why changing one's own attitudes is not complementary to, rather than in conflict with, changing the world itself. And, third, pragmatically, no one can doubt that changing the world itself will take (a lot of) time. At least while this effort is ongoing, everyone individually simply has to cope with trying to be an ethical agent in an unjust world.

However, even the weak interpretation of the world-first strategy is, on its own, incomplete. The worry is that it does not speak to ethical concerns about personal responsibility for implicit bias. As several of the chapters of this volume suggest, it is a difficult and important question whether, and how, individuals ought to be held responsible for behavioral expressions of implicit bias. This question is no less pressing when directed toward oneself. What are *my* responsibilities, given the evidence that I likely hold and act upon implicit biases? On the strongest presentation of the worry, political activism can serve as a source of self-deception, leading one to think that one has discharged one's ethical obligations by "fighting the good (political) fight."[13] It is hard to see how the world-first strategy speaks to these concerns. Perhaps it is complementary with them, but as an expression of what the ethics of implicit bias is all about, the world-first strategy is incomplete.

[12] For review, see Dasgupta (2013).

[13] My point is not that any of the defenders of the world-first strategy are themselves self-deceived. The novelist Tom Robbins (2010) expressed my worry nicely, if not a bit snarkily, when he wrote in *Still Life with Woodpecker*: "Political activism is seductive because it seems to offer the possibility that one can improve society, make things better, without going through the personal ordeal of rearranging one's perceptions and transforming one's self."

3.3 The seek/avoid strategy

A second kind of outwardly-focused ethics of implicit bias can be repurposed from the philosophical reception of the "situationist" literature in social psychology. Critics of virtue ethics like Gilbert Harman (1999) and John Doris (2002) have made clear how unexpected and seemingly trivial features of the situations that we are in can strongly affect our behavior and attitudes. It is not a stretch to think of context cues that affect implicit attitudes as examples of the kinds of situational influences on behavior with which these authors are concerned. There is a sense in which Gawronski and Cesario's background colors in IAT trials are analogous to the oft-cited Isen and Levin (1972) dime in a payphone. Similarly, the conceptual and motivational elements of context that affect implicit attitudes are analogous in many respects to other central situationist examples, such as Darley and Batson (1973) and Milgram (1974/2009).[14]

Situationists tend to be wary of agent-centered ethical ideals. Instead, some situationists argue that the best avenue to ethical action is to focus our attention on the situations we are in and the effects those situations have on us. An obvious way to do so is to seek out situations that are likely to promote wanted attitudes and behavior and avoid situations that are likely to compromise wanted attitudes and behavior. Hagop Sarkissian (2010) calls this the "seek/avoid" strategy, which I here repurpose in the related context of combating implicit bias. Harman (2003: 91) articulates the seek/avoid strategy clearly: "If you are trying not to give into temptation to drink alcohol, to smoke, or to eat caloric food, the best advice is not to try to develop 'will-power' or 'self-control'. Instead, it is best to head [*sic*] the situationist slogan, 'People! Places! Things!' Don't go to places where people drink! Do not carry cigarettes or a lighter and avoid people who smoke! Stay out of the kitchen!"

Unfortunately, the seek/avoid strategy is often hamstrung in day-to-day life, as Sarkissian makes clear. He offers four reasons (2010: 5). First, one has to know which situations to avoid ahead of time, and many situations are neither good nor bad *simpliciter* such that one can know to seek or avoid them ahead of time. Second, some problematic situations are practically unavoidable. Third, there are times when one's ethical commitments themselves require one to enter compromising situations. And fourth, the problematic variables inherent in ordinary situations are so finely individuated that it is hard to know how agents could ever

[14] Nothing I say in this chapter should be construed as support or defense of the specific situationist critique of virtue ethics. I do not know whether whatever traits people have substantiate the moral psychology required by virtue ethics.

discriminate between them. Consider these arguments cashed out in terms of deciding whether to go to a party while trying to quit smoking. First, you may not know whether people will be smoking at the party; second, some parties are more or less obligatory (e.g. office parties); third, you may have overriding reasons to go to the party, such as talking to a friend who is going through a difficult divorce; and fourth, any number of other situational influences present at the party might complicate the way you have conceptualized it (perhaps the presence at the party of a colleague recovering from lung cancer means that this is a situation that would in fact help you to quit smoking).

Each of these concerns applies to the use of the seek/avoid strategy to combat implicit bias. I had the following experience. Having spent the day hearing talks *at a workshop on implicit bias*, I was excited to unwind over dinner with my colleagues. Unbeknownst to the conference organizers, the meal at the chosen restaurant included a belly-dancing performance. This particular performance struck me (and others, I think) as uncomfortably suffused with familiar and problematic gender associations. Just imagine the optics: in a restaurant full of buttoned-up academics, a scantily dressed woman circles the tables, symbolically prostrating herself in front of people who pay her some—but not much—attention while they eat. This would seem to be a good situation to avoid if one is trying to combat implicit associations between the concept of "sexual object" and women. However, the seek/avoid strategy would be little help here. First, no one could have reasonably known ahead of time that this situation was one to avoid. Second, there was some sense of professional obligation to attend the conference dinner, so even if one wanted to avoid the situation, doing so would have come at some cost. Third, for some of the workshop participants, this was not even a situation to avoid. Rather than ignore the dancer, they took this as an opportunity to express their ethical commitments by showing solidarity with a hard-working woman. And so they praised her great skill and—upon being invited—climbed onto the dinner table and danced with her! I took this to be the enactment of an ethical ideal that required entering an otherwise potentially compromising situation (i.e. an illustration of Sarkissian's third point, above). Finally, these colleagues who showed solidarity with the dancer changed the meaning of the experience, for me at least. Rather than see it as a straightforwardly compromising situation, I now think of it as an illustration of how to be creatively ethical. This illustrates Sarkissian's fourth point: the variables that determine whether a situation is likely to compromise or promote our ethical ends are extremely hard to individuate and identify.

4 A Contextualist Approach

My proposal is modest. It is a reframing of the way we think about the ethics of implicit bias, and it borrows from each of the proposals I discussed previously. When conceptualizing the fight against implicit bias, our proximal focus should not be on harmonizing our internal states alone, nor should it be on changing the world in a broad sense, nor should it be on seeking out the right kinds of situations and avoiding the wrong ones as such. What I propose instead is a contextualist approach that blends all three from the get-go. It focuses on precisely those nodes at which our attitudes are affected by features of the ambient environment, and the ambient environment is in turn shaped by our attitudes and behavior.[15] I will give three examples of what I mean, and in each case I will try to clarify why I take the example to illustrate a contextualist approach to the ethics of implicit bias.

4.1 Physical context cues and desirable renewal effects

In the paper discussed previously, Gawronski and Cesario (2013) suggest that physical context cues can play an important role in the regulation of implicit attitudes. While the literature they discuss emphasizes the return of undesirable, stereotype-consistent attitudes in ABA, AAB, and ABC patterns, they also discuss patterns of context-change in which participants' ultimate evaluations of targets reflect the counterconditioning information they learned in the second block of training. These are the ABB and AAA patterns. The practical upshot of this, Gawronski and Cesario suggest, is that one ought to learn counterstereotyping information in the same context in which one aims to be unbiased (ABB and AAA renewal). And what counts as the "same" context can be empirically specified. Gawronski and colleagues (ms.) show that renewal effects are more responsive to the perceptual similarity of contexts than they are to conceptual identity or equivalence. So it is better, for example, to learn counterstereotyping information in contexts that look like one's familiar environs than it is to learn them in contexts that one recognizes to be conceptually equivalent. It may matter less that a debiasing intervention aimed at classroom interactions is administered in another "classroom," for example, than that it is administered in another room that is painted the same color as one's usual classroom.[16] Finally, Gawronski and

[15] This is, of course, not tantamount to a metaphysical claim about the boundaries between agents' minds and the outer world.

[16] Of course, one cannot always make the relevant predictions, such as the wall-color of the classrooms. An anonymous reviewer raises the worry that this means that contextualism will fall prey to the same worries I raised about the seek/avoid strategy (Section 3.3). In some cases this is

Cesario suggest that if it is not possible to learn counterstereotyping interventions in contexts the same as or similar to those in which one aims to be unbiased, one ought to learn counterstereotyping interventions across a variety of contexts. This is because both ABA and ABC renewal are weaker when counterattitudinal information is presented across a variety of contexts, rather than just one. The reason for this is thought to be that fewer contextual cues are incorporated into the agent's representation of the counterattitudinal information when the "B" context is varied. A greater variety of contexts signals to the agent that the counterattitudinal information generalizes to novel contexts.

These are new findings, in need of further consideration, but they are promising. I take them to be contextualist in nature because they are focused on the fine-grained stimuli that act as occasion-setters for wanted behavior.[17] Importantly, a strategy like this is not predicted by the foregoing ways of thinking about the ethics of implicit bias. The importance of attending to the color of the room in which one practices a debiasing procedure is not clearly predicted by the ethics of internal harmony, since the color of the room is a seemingly random feature of the environment from the perspective of one's reflective goals. Nor is this strategy clearly predicted by the world-first or seek/avoid strategies. It is by no means a form of revolutionary political activism, in which we reshape the social, institutional, and economic forces that perpetuate inequality. And by changing the color of the room, one is neither seeking out nor avoiding any particular kind of

true, since we simply cannot predict the future. In addition, it may not be feasible to fix the relevant context cues in the right way, just as it is not always practically feasible to avoid compromising situations. The difference between contexts, as I discuss them here, and situations, in the sense of the seek/avoid strategy, is one of scale. The seek/avoid strategy recommends seeking out good situations and avoiding bad situations wholesale. Do not go to the bar, for example, if you are trying not to drink. Manipulating physical context cues such as the color of the room in which one practices a debiasing procedure, by contrast, is a way of re-engineering small features of the situations in which we know we will be. To put that another way: while we cannot always predict what the situation will be like, we can predict which contextual elements of familiar situations will have wanted and unwanted effects on our behavior. This is precisely the kind of empirical prediction Gawronski and colleagues' research investigates. Future research might investigate other elements of context that might moderate the expression of implicit biases. Air temperature? Ambient noise? Moreover, how do these features of the context interact? Would a noticeably warm classroom elicit renewal effects despite similarities in the color of the walls? Would renewal effects be stronger if the room has the same color walls and is also a similar temperature to the rooms in which one ultimately interacts outside the lab?

[17] An anonymous reviewer asks how fine-grained these contextual features can be before they become impossible to utilize outside the lab. Not very. We cannot often control the color of walls, air temperature, ambient noise, and so on. But the example I am discussing involves a debiasing procedure performed in the lab, which then has lasting effects outside the lab (hopefully). Certainly, these features of the context can be carefully controlled in the lab.

situation. Rather, one is seeking out small elements of situations and thereby changing the significance of the situation itself.

Promoting desirable renewal effects in this way is a relatively minor tool, all things considered, but it is one best conceived through a proximal focus on one's context as a node of exchange between agents and their ambient environment.

4.2 Behavioral context cues and if–then planning

Implementation intentions are "if–then" plans that appear to be remarkably effective for promoting a wide range of goals. An implementation intention specifies a goal-directed response that an individual plans to perform on encountering an anticipated cue. In the usual experimental scenario, participants who hold a goal, "I want to X!" (e.g. "I want to eat healthy!") are asked to supplement their goal with a plan of the form "And if I encounter opportunity Y, then I will perform goal-directed response Z!" (e.g. "And if I get take-out tonight, then I will order something with lots of vegetables!"). Forming a plan to implement one's goal in this specific conditional format significantly improves self-regulation in a wide variety of domains, including (to name just of few of what is a long, long list) dieting, exercising, recycling, restraining impulses, maintaining focus (e.g. in athletics), avoiding binge drinking, and performing well on memory, arithmetic, and Stroop tasks (Gollwitzer and Sheeran, 2006). If–then planning has also been shown to be effective in the regulation of biased implicit attitudes.

For example, if–then planning has been shown to be effective in reducing bias on IAT scores (Webb et al., 2010), a weapons identification task (Stewart and Payne, 2008), and a shooter-bias task (Mendoza et al., 2010). Saaid Mendoza and colleagues, for example, instructed participants in an implementation intention condition to adopt the plan, "and if I see a gun, then I will shoot!" A simple plan like this is thought to work by increasing the accessibility of cues relevant to a particular goal and automatizing the intended behavioral response. Here the relevant cue—"gun"—is made more accessible—and the intended response—to shoot when one sees a gun—is made more automatic. Peter Gollwitzer and colleagues (2008) explain this cue-response link in associative terms. They write: "...an implementation intention produces automaticity immediately through the willful act of creating an association between the critical situation and the goal-directed response" (326).

It is this link between the critical situation and one's behavioral response that is crucial for illustrating why if–then planning is a tool of the kind of contextualist ethics I am recommending. The heavy-lifting in if–then planning is done by the specification of the situational cue to which one plans to respond (i.e. whatever follows the premise in the "if" clause). Indeed, the effectiveness of this form of planning is strongly moderated by the accessibility of the specified cue, as well as the strength of association between the cue and the behavioral response (Webb and Sheeran, 2008). In order to attain one's egalitarian goals, then, one can focus one's attention outward, away from oneself, toward the context cues that act as instigators for goal-consistent behaviors. Gollwitzer (1993: 173) expresses it this way: "...by forming implementation intentions people pass the control of their behavior on to the environment." This formulation may be too strong, if passing the control of one's behavior to the environment is thought to entail a loss of agency. But I suspect that what Gollwitzer means is that by adopting an implementation intention one automatizes one's goal-striving, and one does so by opening oneself to the effects of particular contexts on one's attitudes and behavior. On this reading, if–then planning represents not just a valuable tool for self-regulation as such, but also represents an example of self-regulation via the strategic arrangement of one's behavior-inducing ambient relations.

As in the previous example, if–then planning is not wholly divorced from the effort to be more internally harmonious, to change the world, or to seek out good situations and avoid bad ones. But what makes if–then planning powerful is not predicted by any of these approaches. It can help to make one more internally harmonious, but only by directing one's attention away from one's internal states. Perhaps if–then planning has the power to change the world, but only by changing the behavior of individuals, one by one. And if–then planning does not recommend seeking out or avoiding situations wholesale, but rather, attending to the crucial features of those situations that promote desirable behavior. In other words, it recommends attending to context.

4.3 Social context cues and reciprocal bootstrapping

A final example is the combating of implicit biases by conceptualizing one's own behavior as an element of *others'* context. Sarkissian advances this strategy in response to the situationist critique of virtue ethics. He writes (2010: 12; emphasis in original):

> We hardly notice it, but oftentimes a kind smile from a friend, a playful wink from a stranger, or a meaningful handshake from a supportive colleague can completely change our attitudes. Such minor acts can have great effects. If we mind them, we can foster a form of *ethical bootstrapping*—that is, we can prompt or lift one another toward our joint moral ends. If situationism is true, then whether any individual will be able to meet her ethical aims on any particular occasion will hinge on the actions and manners of others in her presence, which in turn will hinge on her own. In being mindful of the interconnectedness of our behavior, we not only affect how others react

> to us, but also thereby affect the kinds of reactions we face with in turn. The bootstrapping is mutual.[18]

Seemingly minor things that we do in the presence of others, in other words, help to form the context that shapes how others behave, which in turn affects us. Sarkissian takes advice from Confucian ethics in determining which "minor things" we can control, which in turn can have ethical bootstrapping effects. These include mannerisms, tone of voice, and posture, each of which is a source of "*de*," or moral charisma (Sarkissian, 2010: 9). Empirical literature also supports Sarkissian's point; he cites literature showing that smiling and handshaking increase trust and cooperation between strangers (Scharlemann et al., 2001; Manzini et al., 2009).

Shaping one's interpersonal context by attending to mannerisms, tone of voice, posture, and so on, is particularly valuable in the case of implicit bias because these seemingly minor behaviors are precisely the medium through which implicit bias is often expressed. Such so-called "micro-expressions" of prejudice involve, for example, making more eye contact with white colleagues than with black colleagues during a meeting, or referring to male scholars by their last names and female scholars by their first names.[19]

Attending to these micro-expressions of prejudice is certainly in keeping with the ethics of internal harmony, as the end result is hopefully the harmonization of one's internal states. But, again, the proximal goal motivating one's attention in this case is external to the agent. The (hopeful) result as well is not just the harmonization of one's own internal states, but the interpersonal harmonization of one's attitudes and behavior with those of others. Reciprocal bootstrapping in this sense aims to create a harmonious interpersonal atmosphere, so to speak. The regulation of one's own internal states can even be seen as a felicitous byproduct of the aim to create this mutually supportive atmosphere. One can conceptualize this changed atmosphere as a changed world, but not in the institutional sense meant by the world-first critique. Similarly, by adopting this form of contextualism, one certainly minds the effects of the situation on one's behavior, but not simply by seeking out and avoiding particular situations. Rather, one's behaviors help to constitute the situations themselves, as they bear on others, and ultimately on oneself too.

[18] While I endorse giving others kind smiles and meaningful handshakes, I think one should probably avoid winking at strangers.

[19] On bias and microbehavior, see Brennan (2013, this volume); Cortina (2008); Cortina et al. (2011); Dovidio et al. (2002); Olberding (2014); Valian (1998, 2005).

5 Conclusion

Implicit biases are not "directly" activated or expressed in behavior by mere cultural knowledge of stereotypes. Rather, they are highly context-dependent. This fact has ethical ramifications. In particular, it points to the need to amend our usual ways of conceptualizing the ethics of implicit bias with an outward-focused contextualist ethics, according to which agents and their environments are deeply connected. A contextualist ethics of implicit bias focuses on putting oneself into the right relationship with one's context and thereby helping to create the kind of environment that promotes ethical thought and action. Doing so appropriately incorporates the metaphysical complexities of implicit social cognition into our ethical responses to it.

Acknowledgments

I am very grateful for feedback on this chapter from Asia Ferrin, Bryce Huebner, Daniel Kelly, Alex Madva, Kenny Marotta, Jennifer Saul, Natalia Washington, and two anonymous referees for Oxford University Press. Versions of this chapter were presented at the 2013 meeting of the Society for Philosophy and Psychology and the 2013 meeting of the Eastern Division Meeting of the American Philosophical Association. I benefited from many helpful comments and questions at these meetings.

References

Anderson, E. (2010). *The Imperative of Integration.* Princeton, NJ: Princeton University Press.

Barden, J., Maddux, W., Petty, R., and Brewer, M. (2004). "Contextual moderation of racial bias: The impact of social roles on controlled and automatically activated attitudes." *Journal of Personality and Social Psychology* 87(1): 5–22.

Bargh, J. A., Chen, M., and Burrows, L. (1996). "Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action." *Journal of Personality and Social Psychology* 71: 230–44.

Brennan, S. (2013). "Rethinking the moral significance of micro-inequities: The case of women in philosophy." In Jenkins, F. and Hutchinson, K. (eds.), *Women in Philosophy: What Needs to Change?* Oxford: Oxford University Press: 180–96.

Brennan, S. (this volume). "The moral status of micro-inequities: In favor of institutional solutions."

Cortina, L. M. (2008). "Unseen injustice: Incivility as modern discrimination in organizations." *Academy of Management Review* 33: 55–75.

Cortina, L. M., Kabat Farr, D., Leskinen, E., Huerta, M., and Magley, V. J. (2011). "Selective incivility as modern discrimination in organizations: Evidence and impact. *Journal of Management* 39(6): 1579–605.

Darley, J. and Batson, C. (1973). "From Jerusalem to Jericho: A study of situational and dispositional variables in helping behavior." *Journal of Personality and Social Psychology* 27: 100–8.

Dasgupta, N. (2013). "Implicit attitudes and beliefs adapt to situations: A decade of research on the malleability of implicit prejudice, stereotypes, and the self-concept." *Advances in Experimental Social Psychology* 47: 233–79.

Dasgupta, N., DeSteno, D., Williams, L. A., and Hunsinger, M. (2009). "Fanning the flames of prejudice: The influence of specific incidental emotions on implicit prejudice." *Emotion* 9(4): 585–91.

Devine, P. G. (1989). "Stereotypes and prejudice: Their automatic and controlled components." *Journal of Personality and Social Psychology* 56: 5–18.

Dijksterhuis, A. and Bargh, J. A. (2001). "The perception-behavior expressway: Automatic effects of social perception on social behavior." *Advances in Experimental Social Psychology* 33: 1–40.

Dixon, J., Levine, M., Reicher, S., and Durrheim, K. (2012). "Beyond prejudice: are negative evaluations the problem and is getting us to like one another more the solution?" *Behavioral and Brain Sciences* 35(6): 411–25.

Doris, J. (2002). *Lack of Character: Personality and Moral Behavior*. Cambridge: Cambridge University Press.

Dovidio, J. F. and Gaertner, S. L. (2000). "Aversive racism and selection decisions." *Psychological Science* 11: 319–23.

Dovidio, J. F. and Gaertner, S. L. (2004). "Aversive racism." In Zannam M. P. (ed.), *Advances in Experimental Social Psychology*, vol. 36. San Diego, CA: Academic Press: 1–51.

Dovidio, J. F., Kawakami, K., and Gaertner, S. L. (2002). "Implicit and explicit prejudice and interracial interaction." *Journal of Personality and Social Psychology* 82: 62–8.

Gawronski, B. and Bodenhausen, G. V. (2006). "Associative and propositional processes in evaluation: Conceptual, empirical, and metatheoretical issues: Reply to Albarracín, Hart, and McCulloch (2006), Kruglanski and Dechesne (2006), and Petty and Briñol, (2006)." *Psychological Bulletin* 132(5): 745–50.

Gawronski, B. and Cesario, J. (2013). "Of mice and men: What animal research can tell us about context effects on automatic response in humans." *Personality and Social Psychology Review* 17(2): 187–215.

Gawronski, B., Rydell, R. J., Vervliet, B., and de Houwer, J. (2010). "Generalization versus contextualization in automatic evaluation." *Journal of Experimental Psychology* 139(4): 683–701.

Gawronski, B., Rydell, R. J., Ye, Y., and De Houwer, J. (ms.). "Contextualized representation."

Gawronski, B. and Sritharan, R. (2010). "Formation, change, and contextualization of mental associations: Determinants and principles of variations in implicit measures." In Gawronski, B. and Payne, B. K. (eds.), *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications*. New York, NY: Guilford Press.

Gendler, T. S. (2008a). "Alief and belief." *The Journal of Philosophy* 105(10): 634–63.

Gendler, T. S. (2008b). "Alief in action (and reaction)." *Mind and Language* 23(5): 552–85.

Gollwitzer, P. (1993). "Goal achievement: The role of intentions." *European Review of Social Psychology* 4: 141–85.

Gollwitzer, P., Parks-Stamm, E., Jaudas, A., and Sheeran, P. (2008). "Flexible tenacity in goal pursuit." In Shah, J. and Gardner, W. (eds.), *Handbook of Motivation Science*. New York, NY: Guilford Press: 325–41.

Gollwitzer, P. and Sheeran, P. (2006). "Implementation intentions and goal achievement: A meta-analysis of effects and processes." In Zannam M. P. (ed.), *Advances in Experimental Social Psychology*. San Diego, CA: Academic Press: 69–119.

Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. K. (1998). "Measuring individual differences in implicit cognition: The implicit association test." *Journal of Personality and Social Psychology* 74: 1464–80.

Guinote, A., Guillermo, B. W., and Martellotta, C. (2010). "Social power increases implicit prejudice." *Journal of Experimental Social Psychology* 46: 299–307.

Harman, G. (1999). "Moral philosophy meets social psychology: Virtue ethics and the fundamental attribution error." *Proceedings of the Aristotelian Society* 99: 315–31.

Harman, G. (2003). "No character or personality." *Business Ethics Quarterly* 13(1): 87–94.

Haslanger, S. (2015). "Social structure, narrative and explanation." *Canadian Journal of Philosophy*. DOI: 10.1080/00455091.2015.1019176.

Holroyd, J. and Sweetman, J. (this volume). "The heterogeneity of implicit bias."

Huebner, B. (2009). "Trouble with stereotypes for spinozan minds." *Philosophy of the Social Sciences* 39: 63–92.

Huebner, B. (this volume). "Implicit bias, reinforcement learning, and scaffolded moral cognition."

Isen, A. and Levin, P. (1972). "Effect of feeling good on helping: Cookies and kindness. *Journal of Personality and Social Psychology* 21(3): 384–8.

Jacobson, A. (this volume). "Reducing racial bias: Attitudinal and institutional change."

Kelly, D. and Roedder, E. (2008). "Racial cognition and the ethics of implicit bias." *Philosophy Compass* 3(3): 522–40. doi:10.1111/j.1747-9991.2008.00138.x.

Kubota, J. and Ito, T. (2014). "The role of expression and race in weapons identification." *Emotion* 14(6): 1115–24.

Machery, E. (this volume). "De-Freuding implicit attitudes."

Machery, E., Faucher, L., and Kelly, D. (2010). "On the alleged inadequacies of psychological explanations of racism." *The Monist* 93(2): 228–54.

Madva, A. (2012). "The hidden mechanisms of prejudice: Implicit bias and interpersonal fluency. PhD dissertation.

Madva, A. (ms.). "Biased against de-biasing: On the role of (institutionally sponsored) self-transformation in the struggle Against Prejudice."

Madva, A. and Brownstein, M. (ms.). "The blurry boundary between stereotyping and evaluation in implicit cognition."

Manzini, P., Sadrieh, A., and Vriend, N. (2009). "On smiles, winks and handshakes as coordination devices." *The Economic Journal* 119: 537, 826–54.

Mendoza, S. A., Gollwitzer, P. M., and Amodio, D. M. (2010). "Reducing the expression of implicit stereotypes: Reflexive control through implementation intentions." *Personality and Social Psychology Bulletin* 36(4): 512–23.

Milgram, S. (1974/2009). *Obedience to Authority.* New York, NY: Harper and Row.

Mitchell, J. P., Nosek, B. A., and Banaji, M. R. (2003). "Contextual variations in implicit evaluation." *Journal of Experimental Psychology: General* 132: 455–69.

Neal, D. T., Wood, W., Wu, M., and Kurlander, D. (2011). "The pull of the past: When do habits persist despite conflict with motives?" *Personality and Social Psychology Bulletin* 37: 1–10. doi: 10.1177/0146167211419863.

Nosek, B. A. and Hansen, J. J. (2008). "The associations in our heads belong to us: Searching for attitudes and knowledge in implicit evaluation." *Cognition and Emotion* 22(4): 553–94.

Olberding, A. (2014). "Subclinical bias, manners, and moral harm." *Hypatia* 29(2): 287–302.

Payne, B. K. and Cameron, C. D. (2010). "Divided minds, divided morals: How implicit social cognition underpins and undermines our sense of justice. In Gawronski, B. and Payne, B. K. (eds.), *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications*. New York, NY: Guilford Press: 1–18.

Payne, B., Cheng, C. M., Govorun, O., and Stewart, B. (2005). "An inkblot for attitudes: Affect misattribution as implicit measurement." *Journal of Personality and Social Psychology* 89: 277–93.

Richeson, J. A. and Ambady, N. (2001). "Who's in charge? Effects of situational roles on automatic gender bias." *Sex Roles* 44: 493–512.

Robbins, T. (1980). *Still Life with Woodpecker*. New York: Bantam Books.

Rydell, R. J. and Gawronski, B. (2009). "I like you, I like you not: Understanding the formation of context-dependent automatic attitudes." *Cognition and Emotion* 23: 1118–52.

Sarkissian, H. (2010). "Minor tweaks, major payoffs: The problems and promise of situationism in moral philosophy." *Philosophers' Imprint* 10: 9, 1–15.

Schaller, M., Park, J. J., and Mueller, A. (2003). "Fear of the dark: Interactive effects of beliefs about danger and ambient darkness on ethnic stereotypes." *Personality and Social Psychology Bulletin* 29: 637–49.

Scharlemann, J., Eckel, C., Kacelnik, A., and Wilson, R. (2001). "The value of a smile: Game theory with a human face." *Journal of Economic Psychology* 22: 617–40.

Schmajuk, N. A. and Holland, P. C. (1998). *Occasion Setting: Associative Learning and Cognition in Animals*. Washington, DC: American Psychological Association.

Stewart, B. and Payne, K. (2008). "Bringing automatic stereotyping under control: Implementation intentions as efficient means of thought control." *Personality and Social Psychology Bulletin* 34(10): 1332–45.

Valian, V. (1998). *Why so Slow? The Advancement of Women*. Cambridge, MA: MIT Press.

Valian, V. (2005). "Beyond gender schemas: Improving the advancement of women in academia." *Hypatia* 20: 198–213.

Webb, T. and Sheeran, P. (2008). "Mechanisms of implementation intention effects: The role of goal intentions, self-efficacy, and accessibility of plan components." *British Journal of Social Psychology* 47: 373–9.

Webb, T., Sheeran, P., and Pepper, A. (2010). "Gaining control over responses to implicit attitude tests: Implementation intentions engender fast responses on attitude-incongruent trials. *British Journal of Social Psychology* 51(1): 13–32. doi:10.1348/014466610X532192.

3.3

The Moral Status of Micro-Inequities

In Favor of Institutional Solutions

Samantha Brennan

1 Introduction

Let me begin this chapter with two competing aphorisms, both of which are meant to capture some truth about our approach to life and its problems. My grandmother was known for giving, if not herself actually following, the pithy British financial advice: "Watch out for the pennies and the pounds will look after themselves." On this way of thinking, small stuff adds up. You might not care about pennies. By themselves they are too small to be significant, but they add up to pounds, and you do care about pounds and so you ought to care about pennies. Today, I am more likely to hear: "Don't sweat the small stuff." On this view, small stuff does not matter so much and we should focus instead on the big picture. Details bog us down, and we lose sight of what really matters. We should not lose sight of the forest for the trees. In this chapter I want to argue, in the context of discussions about micro-inequities and their moral importance, that both of these sayings get something right.

This chapter is about micro-inequities and their connection to the problem of implicit bias. It begins by defining micro-inequities, and goes on to discuss what makes them wrong and what solutions might be appropriate given the institutional context in which they occur.

The moral problem posed by micro-inequities is connected to three areas of scholarship in philosophy in which I have an interest.

First, there is an emerging area of scholarship in philosophy on the problem of implicit bias of which this volume is part. Jennifer Saul's Implicit Bias and

Philosophy International Research Project at the University of Sheffield describes the project this way: "Unconscious biases against members of stigmatized groups have been studied by psychologists for decades, but only recently have philosophers explored this phenomenon. This project brings researchers from both fields together with policy professionals to work through the implications."

Second, I have an ongoing research interest in problems about moral aggregation, which concerns the addition of goods and bads (on the value side of moral philosophy) or the addition of rights and wrongs (on the right action side of moral philosophy). For example, in value theory we might ask how to compare the disvalue of many people suffering from the common cold to the disvalue of a much smaller number of people experiencing an early death. This is not just a philosopher's idle thought experiment. Such problems of moral aggregation matter, for example, in the context of decisions about directing funding for medical research. When thinking about right action, we might also ask whether it is worse morally for one person to steal $100 or for a hundred people to steal $1 each.

In my development of the "threshold account of rights," according to which rights are overiddable on the basis of the good results that can be brought about by doing so, I have had to struggle with questions about the aggregation of small harms. If your account of rights is an account of absolute rights, then it does not matter at all what else hangs in the balance. No amount of good at stake can justify infringing an absolute right. The situation is more complicated when you believe that there are circumstances which justify infringing a right. Is it just a matter of the total amount of good that hangs in the balance, or are there restrictions on how this total is met? One counterexample against overridable rights plays on the aggregation of small harms. Here is the short version of the "no lives for headaches" tradeoff with which you are probably familiar: If rights not to be killed are overridable on the basis of aggregated harms to others, then there is some number of headaches which would be great enough to add up to justify taking a life.[1] That conclusion rests on the mistaken assumption that there are no limits on moral aggregation, and so rights with thresholds can be overridden in the face of a great many small benefits that can be brought about by infringing the right or small harms that can be avoided by doing so. I have also written about aggregation in the context of debates about inequality.[2]

Third, there is philosophy's recent willingness to subject our own practises related to diversity and exclusion to philosophical scrutiny. As a Department Chair, a member of the Women in Philosophy Taskforce, a member of the

[1] Brennan (1995); Brennan (1997).

[2] Brennan (2009).

Canadian Philosophical Association's Equity Committee, as a blogger with the Feminist Philosophers blog, and simply as a feminist and a concerned member of our profession, I have spent time thinking and writing about how we might do better. Drawing on these experiences, I have contributed to the growing literature within philosophy about our discipline's gender problem.[3]

How are micro-inequities connected to the problem of implicit bias? An emerging story about the persistence of workplace inequality—in the absence of formal barriers to entry and progress for women, minorities, and disabled persons—looks to the twin causes of implicit bias and micro-inequities. The Barnard Report on Women, Work, and the Academy describes these causes of inequality in the academy in these terms: "The first is that biases operating below the threshold of deliberate consciousness, biases in interaction that are unrecognized and unintended, can systematically put women and minorities at a disadvantage. Second, although individual instances of these 'micro-inequities' may seem trivial, their cumulative effects can account for large-scale differences in outcome; those who benefit from greater opportunity and a reinforcing environment find their advantages compounded, while deficits of support and recognition ramify for those who are comparatively disadvantaged" (MIT, 1999: 10).

What can philosophers contribute to our understandings of micro-inequities and their moral importance? An earlier paper set out to examine micro-inequities in the context of women's careers in the academic discipline of philosophy.[4] It offered a short philosophical analysis of micro-inequities and looked at some explanations of why moral philosophy has struggled with the problem of small-harms. There I argued that we need to rethink the moral significance of micro-inequities. While the initial idea of a micro-inequity is easy to understand—specific examples, useful analogies, and colorful metaphors abound—what seems to be missing is careful moral analysis of the micro-inequities. This is the contribution which moral philosophers can make to the discussion of micro-inequities and which this chapter sets out to make.

That said, there are aspects of this problem which this chapter will not be addressing. Specifically, I will not be addressing the mechanisms by which micro-inequities add up to larger harms. This chapter will assume that this happens in the ways that workplace sociologists claim it does, but will leave the details of that account to people with other disciplinary skills.[5] There are also aspects of the problem relevant to moral philosophy that I will not address. Specifically, I will not be discussing praise and blame and moral responsibility directly, but what

[3] Brennan (2013); Brennan and Corless (2009).

[4] Brennan (2013).

[5] See, for example, Virginia Valian's (1999) discussion of "accumulated advantages."

I will discuss may be relevant to questions about moral responsibility. There are hard and interesting questions here about collective responsibility and workplace climate which I do not have the expertise to address and which fall outside the scope of this essay.[6]

There are also some claims I am *not* making, and I want to be clear about these at the outset. It is important to note that I am not claiming that micro-inequities are the *only* problem facing women, the disabled, and other minorities in the workplace. There are still persistent big problems that are not in any sense "micro." Sexual harassment, for example, is not a micro-inequity, and it still matters very much. I take seriously Claudia Card's claims about the danger of focusing on equality over evil (though I also think that inequalities are connected to evils in ways that matter).[7] The issue is even worse if we are focusing on micro-inequities over evils.

There are three reasons why I am interested in understanding micro-inequities. First, I do think they are a big part of the story of how racism, sexism, and ableism persist in places in which, on the face of it, there are rules which prohibit bias and in which most of the people would support those rules.

Second, I am also by temperament much more interested in cases of wrongdoing by good people—that is, by people who think they are doing the right thing but who end up with an outcome they do not endorse. For example, consider the problem of justice in the home. It is no surprise that couples who hold particular conservative views about gender roles have an uneven division of household labour. After all, that particular division of work is what they want. It is much more puzzling when opposite couples with egalitarian commitments end up with a traditional division of household work.

Third, I am interested in hard problems in ethics, and the hardest problems in ethics are not necessarily about the cases that involve the most wrong. Let me explain. There are cases in ethics that just are not that difficult. Lying to hurt a friend, or murdering a dog to make your neighbour cry, are both bad acts done for bad reasons, and they bring about terrible results. Their wrongness is overdetermined. They are terribly wrong, but not terribly morally controversial. Unfortunately, much of life is like that. Understanding *why* people do wrong is a tough problem in moral psychology, but it is easy to see *that* these acts are wrong.

2 Defining Micro-Inequities

Micro-inequities are small, unjust inequalities often pointed to as part of the larger story about larger scale inequalities, such as women's unequal place in the

[6] See work by my Western colleague, Tracy Isaacs (2011).

[7] Brennan (2009).

workforce. But one does not find anything close to a precise definition of a micro-inequity in the literature on workplace ethics and equity. What exactly is a micro-inequity? People often contrast inequities with mere inequalities, where the former are taken to be unjust inequalities. The latter term, "inequality," is a normatively neutral term, while "inequity" assumes that there is some injustice involved. An inequity is a harm that derives its wrongness from being an undeserved inequality. Micro-inequities are very small inequities in which the "smallness" is measured by their results. As far as I know, there are inequities and micro-inequities. No one talks about mini-inequities which would be halfway between a full-sized inequity and a micro-inequity. So the term "micro-inequity" is used for any size of inequity which falls shy of a counting as a full-blown inequity on the basis of its size.

If size is the determining feature of a micro-inequity, what is the relevant size? It seems unlikely to me that we will find an answer that works across the board. My general view about the size of harms and their moral relevance is that it depends on what we are comparing them against. The answers to such questions are context dependent rather than a matter of underlying moral metaphysics. Let us look then at the context in which this discussion is taking place. For practical purposes, in the university and workplace context, I suggest that we think of micro-inequities as inequalities that fall beneath the threshold of legislation or formal actionability.[8] Imagine a person walking into a university equity office and presenting one example of unequal treatment. If that unequal treatment had only small effects and there was only that one instance, it would likely be dismissed as not being, on its own, the sort of thing with which the equity office was charged to deal.

The following are some definitions of micro-inequity from the literature on workplace climate:

1. According to Bernice Sandler, "micro-inequity" refers to the ways in which individuals are "either singled out, or overlooked, ignored, or otherwise discounted" based on an unchangeable characteristic such as race or gender. A micro-inequity generally takes the form of a gesture, different kind of language, treatment, or even tone of voice.[9] It is suggested that the perceptions that cause the manifestation of micro-inequities are deeply rooted and unconscious. The cumulative effect of micro-inequities can impair a

[8] Later I suggest that informal actionability—such as "calling out"—might be one response to micro-inequities.

[9] I am not endorsing Sandler's definition. Note that it is so broad as to include things we properly think are macro-inequities.

person's performance in the workplace or classroom, damage self-esteem, and may eventually lead to that person's withdrawal from the situation.

2. Mary Rowe, responsible for introducing the term, defines "micro-inequities" as "apparently small events which are often ephemeral and hard-to-prove, events which are covert, often unintentional, frequently unrecognized by the perpetrator, which occur wherever people are perceived to be 'different'."

Rowe named one of her articles the "Saturn's rings phenomenon" because the planet Saturn is surrounded by rings, which obscure the planet but are made just of tiny bits of ice and sand. Rowe writes that her interest in these phenomena began with an incredible opportunity: "In 1973 I took a job at MIT, working for the then new President and Chancellor. I was charged, among other things, with learning how the workplace could improve with respect to people who were underrepresented at MIT—as examples, men and women of color, white women, and people with disabilities. As an economist I had expected to learn about big issues standing in the way of progress." She writes that she did find a few big issues, but not as many as she expected and not enough to account for the scope of the problems. What struck her instead were all of the "little issues." She writes that little acts of disrespect, and failures in performance feedback, seemed to corrode some professional relationships like bits of sand and ice.

What are some examples of micro-inequities? Philosopher Berit Brogaard gives the following list in her blog post about micro-inequities for Psychology Today:[10]

- Checking emails or texting during a face-to-face conversation.
- Consistently mispronouncing a person's name.
- Interrupting a person mid-sentence.
- Making eye contact only with males while talking to a group containing both males and females.
- Taking more questions from men than women.
- Confusing a person of a certain ethnicity with another person of the same ethnicity.
- Rolling your eyes.
- Sighing loudly.
- Raising your voice, even though the other person has no difficulties hearing you.

[10] See "Micro-Inequities: 40 Years Later." <http://www.psychologytoday.com/blog/the-superhuman-mind/201304/micro-inequities-40-years-later>.

- Mentioning the achievements of some people at a meeting but not others whose achievements are equally relevant.
- Consistently ignoring a person's emails for no good reason.
- Only reading half of a person's email and then asking the person about the content later.
- Making jokes aimed at certain minority groups.
- Being completely unpredictable in your grading of certain people's term papers.
- Issuing invitations that are uncomfortable for certain groups ("Please feel free to bring your wife," "There is a link below to childcare options for female speakers who plan to bring their children," "There will be a belly-dancer at the party," "The Department's annual Christmas party will be held on December 18," "Please bring pork chops to the potluck dinner")[11]

You can also find many more examples at the blog, "What is it like to be a woman in Philosophy," though unfortunately most of those examples are too big to count as micro-inequities.

When I have needed an example of an implicit bias-related micro-inequity that is clearly trivial I have turned to the example of waiting times in coffee shops. This example illustrates both how trivial and how pervasive micro-inequities can be. Economists (and their graduate students) set out to measure how long different people waited once they placed an order for coffee.[12] It turns out women wait, on average, twenty seconds longer than men do for a comparable order. This is not a case of women ordering fancy, high-maintenance beverages. and men ordering plain black coffee. African Americans also waited longer than white Americans, and the less attractive waited longer than the more attractive. In and of themselves the extra time waiting is not much of a harm. Likely it is not even noticed by those preparing the drinks or those waiting. The extra twenty seconds even compounded over days, months, years of ordering coffee in a sexist society is not a big deal. Would you organize a protest? Likely not.

But aggregate these micro-inequities—or, as I have called them elsewhere, everyday inequalities—across all aspect of one's life and they begin to make a difference.

I also think it is important to distinguish micro-inequities from another term that is used in discussions of workplace climate: "micro-aggressions."

[11] Note that most of Brogaard's examples do count as intentional acts. What is not intentional is the harming of the other person. Thanks to an anonymous reviewer for pointing this out.

[12] Myers et al. (2010).

Micro-aggressions are defined as "subtle, verbal, and nonverbal insults directed toward non-whites, often done automatically and unconsciously. They are layered insults based on one's race, gender, class, sexuality, language, immigration status, phenotype, accent, or surname." The term was coined in 1970 by a psychiatrist to describe acts of racism so subtle that neither the "perpetrator" nor the "victim" is even fully conscious of what is happening. "The invisibility of racial microaggressions may be more harmful to people of color than hate crimes or the overt and deliberate acts of White supremacists such as the Klan and Skinheads," writes Derald Wing Sue, in *Racial Microaggressions in Everyday Life: Race, Gender and Sexual Orientation.*

What is the distinction between micro-inequities and micro-aggressions? Note that while neither requires active intention on the part of the person whose actions bring about the micro-inequity or micro-aggression, micro-aggressions are defined primarily in terms of "insults." Not all micro-inequities involve the expression of a view about a particular group or individual. As well, some micro-inequities result from small positive acts unjustly distributed. Again, this shows that not all micro-inequities are micro-aggressions.[13] I find it helpful to think of micro-aggressions as a subclass of micro-inequities. One thing that is clear in the literature on micro-inequities is that they can be brought about by a member of the group in question—i.e. women are as likely as men to perpetuate them against women as are men—and that seems harder to make sense of in the case of micro-aggressions. My worry about the language of "micro-aggression" is that wrongdoing and culpability for the result seem built into the idea of aggression and aggressive behaviour, and that is not the case for micro-inequities. Indeed, the question of wrongdoing and micro-inequities is part of what is at issue here.

3 Useful Metaphors

While coming up with a single definition of "micro-inequity" is not easy, there are many colorful metaphors that capture what is philosophically interesting about them. Here I will mention three of them and say something about the specific aspect of the problem posed by micro-inequities that they help illuminate.

Marilyn Frye's birdcage metaphor helps us understand why focusing on each individual micro-inequity is a mistake. Instead, what is interesting from the perspective of understanding how oppressed groups are treated in the workplace and in universities is the pattern formed by the micro-inequities. Each individual

[13] Thanks to Jennifer Saul for this point.

instance, looked at alone, cannot explain the larger wrongdoing. Marilyn Frye writes:

> Consider a birdcage. If you look very closely at just one wire in the cage, you cannot see the other wires. If your conception of what is before you is determined by this myopic focus, you could look at that one wire, up and down the length of it, and be unable to see why a bird would not just fly around the wire any time it wanted to go somewhere. Furthermore, even if, one day at a time, you myopically inspected each wire, you still could not see why a bird would gave trouble going past the wires to get anywhere. There is no physical property of any one wire, nothing that the closest scrutiny could discover, that will reveal how a bird could be inhibited or harmed by it except in the most accidental way. It is only when you step back, stop looking at the wires one by one, microscopically, and take a macroscopic view of the whole cage, that you can see why the bird does not go anywhere; and then you will see it in a moment. It will require no great subtlety of mental powers. It is perfectly obvious that the bird is surrounded by a network of systematically related barriers, no one of which would be the least hindrance to its flight, but which, by their relations to each other, are as confining as the solid walls of a dungeon. It is now possible to grasp one of the reasons why oppression can be hard to see and recognize: one can study the elements of an oppressive structure with great care and some good will without seeing the structure as a whole, and hence without seeing or being able to understand that one is looking at a cage and that there are people there who are caged, whose motion and mobility are restricted, whose lives are shaped and reduced.[14]

Or consider the metaphor of a "ton of feathers" which is often used to describe the kinds of things that go on in chilly climates.[15] The idea is that having a single feather land on you is harmless, at worst annoying, but a ton of them is deadly.[16] While the birdcage metaphor is useful in helping to show why we do not see micro-inequities, the "ton of feathers" analogy explains why we do not mind or even notice one or two micro-inequities, but large numbers matter.

These metaphors also relate to two different aspects of the invisibility of micro-inequities. Micro-inequities are sometimes said to be invisible because of an epistemological shortcoming on the part of moral agents, as both the perpetrators of micro-inequities and the victims of micro-inequities. We do not see them because our moral sensors are not tuned to see sufficiently small harms. If we looked at the birdcage properly we would notice the way it is functioning and the role it plays in constraining the bird, for example. The feathers analogy is a bit different. One feather landing on your head is no harm at all. It is not that it is a small harm that we do not see. Rather, the size of the feather and the lack of impact it causes means it is invisible as a harm because it is not a harm

[14] Frye (1983). [15] Caplan (1993). [16] Caplan (1993).

at all. I return to these two different ways that micro-inequities can be invisible when I discuss what makes micro-inequities wrong.

4 How are Micro-Inequities Connected to Implicit Bias?

The Barnard Report calls micro-inequities and implicit bias the twin causes of women's inequality in the academy. This way of describing implicit bias and micro-inequities makes it seem as if they were two completely different phenomena. Other places in the literature researchers will run implicit bias together with micro-inequity and you might, after reading, wonder what the difference is between them. It is my view that these categories will often overlap, though they need not. Not every micro-inequity will be the result of implicit bias, and not all cases of implicit bias will result in micro-inequities. There is also an obvious difference between implicit bias and micro-inequities in terms of the focus of evaluation: the latter is concerned with the *cause* of behavior, and the former with its *nature* or *effects*.[17]

For example, a micro-inequity could follow from an intentional act of bias. I could, on the basis of some prejudice I willingly admit, grade some students more harshly than others. For example, a lecturer who repeatedly passes over female students' contributions could be aware of doing so and intend to because of a conviction that the contributions will not be valuable. In this case, each time a female student is not called on a micro-inequity results, but it is not one that results from *implicit* bias. Likewise, the inequalities that result from implicit bias might be large and substantial. They might be not "micro" at all. Consider the case of "shooter bias" in which White police officers and undergraduate students mistakenly shoot unarmed Black suspects more than White suspects on computerized shoot/do not shoot tasks. This can result from implicit bias and the results are very serious indeed.[18]

Why do they usually go together? I have some suggestions. First, it is rare for a person to intend another person a small harm. That is, even if my bias is up front, I usually will not want to wrong someone a little bit.[19] If I am right about this,

[17] Thanks to an anonymous referee for noting this difference.

[18] Correll, Park, Judd, and Wittenbrink (2007).

[19] I am not suggesting that we all follow Machiavelli's advice: "Upon this, one has to remark that men ought either to be well treated or crushed, because they can avenge themselves of lighter injuries, of more serious ones they cannot; therefore the injury that is to be done to a man ought to be of such a kind that one does not stand in fear of revenge" (*The Prince*, 1513). This is often paraphrased as "never do an enemy a small injury."

then intentional harms—harms that come about as a result of explicit bias—are rarely small. Second—and I am less certain about this one—I think we would come to notice our implicit biases if they regularly brought about large differences in treatment. Part of what keeps implicit biases unnoticed is that the resulting harms are quite small. It is only in the aggregate that they are noticed.

5 What Makes Micro-Inequities Wrong?

In canvassing plausible views about the wrongness of micro-inequities there is one view I want to note at the outset, reject, and then move on. I think we ought to reject the view that micro-inequities are not usually wrong because of the previously discussed connection with implicit bias. One answer to the question of micro-inequities and their wrongness is that micro-inequities are not wrong unless they are intentional. Of course, there are some complications here. What does my intention have to be? My view is that an intention to do wrong is not necessary for wrongness. Consider the standard example from utilitarian moral reasoning in which we choose to promote our own good rather than donate money to a worthy charity. I might forgo the Oxfam change box and instead spend my coins on coffee. In such a case I need not intend a child's death, but it still might be wrong to bring about an outcome in which I get a coffee and a child dies.

Let us assume, then, with me, that intention is not required for wrongness. Intention might be required for blameworthiness, but praise and blame are a different matter from right and wrong. We should not assume from the claim that it would be a mistake to blame someone if they intended no harm that what they did was not wrong.

We can then ask, what makes micro-inequities wrong? Or to put the question differently, in what does the wrongness of micro-inequities consist? One obvious answer is that the wrongness of micro-inequities stems from their connection to morally objectionable macro-inequities. But do micro inequities have any moral significance independent of the fact that they are relevant for explaining an objectionable macro-inequity? How might we think about the status of the micro-inequities taken in isolation?

Here are two possible answers:

1. *Threshold Wrongness* Consider a threshold view where the micro-inequities are each themselves, taken individually, too small to count to make a moral difference. We could describe them as causing badness that is beneath the threshold for wrongness. But they cumulatively cause the morally objectionable macro-inequity. So individually they are not wrong, but they can add up to a wrong.

Let me explain in a bit more detail. Suppose that for a given bad effect it needed to be bad to a level of some arbitrary number, say 3, for it to be wrong. And suppose what we had were lots of actions that resulted in 2s. If someone were the recipient of twenty such 2-level actions, that would be bad indeed. A great wrong would have been perpetrated, even though none of the individual actions was wrong.

2. *Strict Additive Wrongness* We could say instead that the micro-inequities are only a little morally objectionable taken in isolation. They are bad and wrong, but not wrong enough to take action against them, as a practical matter. They cumulatively cause the morally objectionable macro-inequity. So individually, they are a little bit wrong, and they add up to a larger wrong.

This view preserves the idea of strict decomposition. The wrong of the whole is a direct additive function of the parts. But there are some other possible ways that this might work.

3. *Organic Whole Wrongness* The wrong of the whole can be greater than the wrong of the parts. Obviously I am borrowing language here from G. E. Moore's discussion of Organic Unities.[20] The idea here is that each small action is wrong, but the wrongness of the collection is worse than the sum of the wrongness of each of the parts. This version preserves monotonicity, and each increase in the small level wrongs leads to an increase in the overall wrongness, but the function that takes us from the micro-inequities to the larger wrong need not be strictly additive.

4. *Not wrong at all, too small* It is also possible that micro-inequities are not wrong at all. They are too small to be wrong, and it is only the macro-inequities that are wrong. If that is right, then we need to provide an account of collective responsibility for the large-scale macro-inequalities.

Now, the problem of small harms is not new. In my earlier paper on micro-inequities I looked at some of the reasons why moral and political philosophers have overlooked them, and then explored some of the work that has been done on small harms, usually from the perspective of environmental harm. Here I will just mention one of them, because it is a work that will be familiar to most people who work in ethics.

Derek Parfit draws our attention to small harms in *Reasons and Persons*,[21] in which he considers a series of mistakes in moral mathematics, including the mistake of ignoring small or imperceptible effects. Even if imperceptible, bad effects with sufficient extent or repetition can be very terrible indeed. Parfit's

[20] See Moore (1903).

[21] Parfit (1984).

examples concern environmental issues such as overfishing and pollution, but his lesson can be just as important for small injuries and insults that are part and parcel of academic life for some people. If we view each act individually, we might miss out on the aggregative effects and on the patterns that are relevant to understanding bias and discrimination.

My position is that the question of whether some benefits and burdens are so small that they do not count morally cannot be answered outside of the context in which they occur, and that we lose sight of morally important factors if we push all the time to see wrongness in its smallest possible units. This is especially true given how much of the rightness/wrongness question here rests on harm, and harm is a very lumpy good. What do I mean by "lumpy good"? Economists use this term to explain cases in which value does not add up and decompose in the units one might expect. Here is a standard example from economics. If I need 100 feet of fence to enclose my backyard because I have a dog I like to let roam freely without risk of escape, then 99 feet of fence is not almost as good as 100 feet. Instead, it is almost as good as no fence at all, though perhaps just a little bit better. Now consider an example from workplace climate. I might merrily endure an ongoing series of small slights from colleagues, until one day someone says something not terribly mean or hurtful, and I am plunged into depression. If it is genuinely true that I suffer no ill effects until one remark tips me into a serious depression, then what I have experienced is a lumpy bad. Stopping one remark earlier and I would have experienced no harm, just like stopping 1 foot of fence shy of enclosing the garden is no good at all. Paying too much attention to that one act of speech misses the point, I think.

There are also further wrinkles and other factors we need to consider when we are thinking about workplace wrongs. It is useful to distinguish between different kinds of wrong. Some micro-inequities are wrong purely because of facts about distribution. In the academy we might think that academic service work is neutral, but if women bear more than our fair share of committee work, the only wrongness comes from the unequal distribution. Other things might be positive—say, for example, research grants—and again, wrongness would stem from an unjust distribution. In other cases, inequality is a smaller part of the story. Sexual harassment is wrong simpliciter. How much of an improvement would it be if it occurred equally between the sexes?

6 Is Focusing on Individual Act Wrongness a Mistake?

While the question of an individual act's wrongness is important, I worry that focusing too much on it may cause us to lose sight of the forest for the trees. My own answer is that sometimes the individual micro-inequities will be wrong, even though we may not be in a position to blame anyone. Other micro-inequities may be neither wrong nor blameworthy. Some might worry that not addressing the wrongness of the individual micro-inequities may seem to be letting people off the hook too easily, but it does allow us to shift our focus to collective solutions to the problem. As a group we have responsibilities for the outcome, and group-based solutions are likely to be much more effective than individual solutions. Finally, some may worry that the focus on micro-inequities lets us all off the hook for the large-scale culpable wrongs that do occur in the academy. There are also important questions about why we have the implicit biases that we do. Are we not, as a society, responsible for our sexist, racist, homophobic, and ableist beliefs even if they are implicit in our thinking?[22] My answer here is that there are questions both larger (for example, the societal beliefs that inform implicit bias) and smaller (for example, individual responsibility), but that the most practical place to address the issues is at the level of the group in which we find ourselves—in the middle, at the department, and at university level.

7 In Favor of Institutional Solutions

Why, then, do philosophers spend most of our time thinking about micro-inequities in the context of wrong actions, rather than in the context of changing the contexts in which they occur so as to make them less likely or to repay those who unfairly bear the costs? I think it is a bit like the analogy of looking for lost keys under a lamplight. We look there, not because that is where the keys fell, but rather because that is where the light is. Moral philosophy has a rich vocabulary about individual moral wrongs, but as Tracy Isaacs notes, we are less able to deal with those wrongs that occur in collective contexts.[23]

I am not saying that the individual wrongs that are associated with micro-inequities are unimportant. I am saying that they are only part of a larger picture, and we get it wrong when we ignore other aspects of the problem. Consider these three different sites of moral inquiry: (1) circumstances under which decisions and choices are made; (2) the acts themselves; and (3) the results. I argue that focusing on (2), the question about the wrongness of the acts themselves, is potentially dangerous for movements interested in social change. It is an important question in moral theory, but it might not be the most important question for us.

[22] See Volume 1: chapters by Frankish (1.1), Huebner (1.2), Machery (1.4), and Mallon (1.5).

[23] Isaacs (2011).

Let us begin by considering an older argument from political philosophy: Robert Nozick's Wilt Chamberlain case.

> Wilt Chamberlain is greatly in demand by basketball teams, being a great gate attraction. (Also suppose contracts run only for a year, with players being free agents.) He signs the following sort of contract with a team: In each home game twenty-five cents from the price of each ticket of admission goes to him. (We ignore the question of whether he is "gouging" the owners, letting them look out for themselves.) . . . Let us suppose that in one season one million persons attend his home games, and Wilt Chamberlain ends up with $250,000, a much larger sum than the average income and larger even than anyone else has. Is he entitled to his income? Is this new distribution D2 unjust? . . .

Put differently: Nozick asks us to imagine a just starting point. Let us suppose we are egalitarians and that the just starting point is an equal distribution of goods. Through a series of very small free exchanges we end up in a situation that is very unequal in terms of wealth. Wilt Chamberlain has a lot more money than everyone else. But if it is wrong that Chamberlain has lots more than the rest of us, how did the wrong come about? Nozick pushes us to either accept that the situation where Wilt is rich and everyone else is not as just, or admit that each act of giving Wilt a quarter to play basketball was morally wrong. Nozick's view is that wrongs cannot be mysterious and that the only way big wrongs can come about is through the aggregation of small wrongs. If none of the acts of giving Wilt a quarter to see him play was wrong, then the outcome cannot be wrong. This relates to the previous discussion, because we can think of the choice to give Chamberlain a quarter as bringing about a micro-inequity. Note that Nozick's answer assumes that Strict Additive Wrongness is correct. We can think of the "quarter for basketball" exchanges as small wrongs which do add up to one larger wrong. That is what Strict Additive Wrongness demands. On Nozick's view, Threshold Wrongness is incorrect. It is mysterious. Big wrongness can only come from small wrongness, on his view. According to Threshold Wrongness, each little act could be inequality-making but not enough to make it wrong. Wrongness is an emergent property. It comes about when the harms and inequalities are large enough.

Let us suppose though that Nozick is right. Let us give him Strict Additive Wrongness. Does that mean that the acts of giving Chamberlain a quarter to watch him play should be forbidden? Not necessarily. Banning the transaction at the level of individual exchanges is not the only possible response. We may have other choices. For example, suppose that banning the transactions and allowing them and fixing the inequalities that result through taxation bring about the same result. For example, it may be that we infringe liberties less when we fix things like Chamberlain's extra income on an annual basis, through taxation, for example.

Let me give you another example of this sort. I favor my own children, both in social ways that benefit them and also in financial ways. This favoritism produces inequalities, which occur both because I have more income to pay for classes and activities that benefit my children, and also because I might have more time to devote to reading to them, for example. But does that mean that the state should interfere with my parenting to produce a more equal outcome? No. Clearly there are less liberty-infringing ways to bring about a more equal outcome. One can tax income, use taxation revenue to develop and promote programs for needy children, and address inequalities through a well-funded system of public education.

In summary, I think there are three reasons not to move too quickly to the level of individual action and its wrongness.

First, there are very hard questions about the facts of how harms and inequalities add up in any given case. Second, the dispute in moral philosophy between Threshold Wrongness and Additive Wrongness will be difficult to resolve. I think Threshold Wrongness is correct, but I take it that many will be unpersuaded. If a lot rode on getting this right, then we ought to persevere, but my next point is that whether or not these small inequality producing acts are wrong will not have much import. That is because, third, there are costs in interfering with individual actions that are hard to bear. These costs are both in terms of implementation and in terms of liberty. It is better in the case of small inequalities to create contexts in which they are less likely to come about.

Let me now turn to some objections. First, does this let those who bring about micro-inequities off the hook too easily? Is there nothing we can do? No. It is not the case that we need do nothing. I have, in fact, three suggestions.

One obvious tool at our disposal is that of blame. Now, not all micro-inequity-producing actions will be blameworthy. As I stated at the outset, I will not outline the conditions that make actions worthy of praise or blame. It might be that the person needs to know they are causing harm, for example. Supposing that these conditions are met, it is still going to be true that they are to blame for a small harm. In a paper that I am writing with Meghan Winsby, a Western PhD student, we are proposing micro-sanctions as an appropriate interpersonal response to those who contribute in a small way to chilly workplace climates.[24] Examples of micro-sanctions include failing to give uptake to a racist joke and correcting a speaker who uses sexist language; Alex Madva and Michael Brownstein have

[24] Brennan and Winsby (in preparation).

proposed “implicit reactive attitudes”—subtle gestures expressing praise/blame—as ways of coping.[25]

Second, while we might use micro-sanctions to indicate our disapproval of those who knowingly bring micro-inequities about, I have suggested micro-affirmations as a way of reaching out to those who suffer from micro-inequities. Micro-affirmations may take the shape of deliberately reaching out to a student, colleague, or coworker who is isolated. One might make a special point of recognizing this person’s contribution in the workplace. The idea is that positive micro-messaging can redress and rebalance the harms caused by micro-inequities. For example, not all celebrations of a person’s research need involve awards, banquets, and trophies. Respectfully engaging colleagues in discussions of their ideas may be more professionally valuable. Mentioning a colleague’s work when relevant, in a way that demonstrates your awareness of it, might also be a micro-affirmation.[26]

Third, there will be obligations that fall on those who bring about micro-inequities once we know about implicit bias and about how unsuccessful individual attempts to “try hard and do better” can be. There is indeed some evidence that mere awareness coupled with a resolution “to be objective” might actually bring about worse results. But some active debiasing programs have been shown to have a positive effect.[27]

A second objection comes from the work of libertarian economist Bryan Caplan,[28] who argues that bias is not just an issue in the workplace for the traditionally disadvantaged groups such as women, racial and ethnic minorities, sexual minorities, and the disabled. It also affects—recall my coffee example—the ugly, the overweight, and the generally unlikable. Caplan continues in a reductio of the argument against aggregation and bias that bias does not end at the door of the academy. Implicit bias and micro-inequities affect almost all the fabric of our personal lives, from dating and friendship, to retail and banking exchanges. Back when we waited for bank tellers to fetch us our cash, no doubt some of us waited longer than others. Should the banks have mailed a waiting bonus to women, to African Americans, and to the less attractive bank customers? Caplan thinks this result is clearly absurd, and so we are wrong to worry about bias and about small harms.

[25] Brownstein and Madva (2012).

[26] For other examples of positive micro-affirmations, see Young (2007).

[27] Uhlmann and Cohen (2007). See also Kenyon (2014) and Kenyon and Beaulac (2013).

[28] Bryan Caplan, personal correspondence.

My response to Caplan is two-fold. In the first instance, I will bite his bullet and agree that we ought to work towards eliminating all sorts of bias. An advantage of institutional solutions such as anonymous grading is that they protect others outside the scope of what we call in Canada, target groups (in Canada, women, disabled, racial minorities, and indigenous persons). In the second, I deny that there are justice-based obligations to reform our personal lives, though there may well be moral reasons for doing so. It may well be morally required to diversify my group of friends and/or not choose to date people on the basis of their attractiveness, but such a moral obligation is hardly the sort of thing that the state can enforce.[29]

I have argued that micro-inequities are morally significant but that the best structural response to them is at the institutional level. I have suggested that we ought to think more about the contexts in which micro-inequities occur and about how to respond to them collectively and focus less of our attention on the wrongness of individual micro-inequities. even though there are interesting moral questions about their status. This is a case in which the morally interesting details drag our attention away from genuine possibilities for real social change.

References

Brennan, S. (1995). “Thresholds for rights.” *The Southern Journal of Philosophy* 33(2): 143–68.

Brennan, S. (1997). “Moral rights and moral math: Three arguments against aggregation.” In Brennan, S., Isaacs, T., and Milde, M. (eds.), *A Question of Values: New Canadian Perspectives in Ethics and Political Philosophy*. Amsterdam: Rodopi Press: 29–38.

Brennan, S. (2009). “Feminist ethics and everyday inequalities.” *Hypatia*: *Oppression and Moral Agency: Essays in Honor of Claudia Card, Special Issue* 24(1): 141–59.

Brennan, S. (2013). “Rethinking the moral significance of micro-inequities: The case of women in philosophy.” In Jenkins, F. and Hutchison, K. (eds.), *Women in Philosophy: What Needs to Change?* Oxford: Oxford University Press.

Brennan, S. and Corless, R. (2009). “Reflections on creating a warmer environment for women in the mathematical sciences and in philosophy.” *Atlantis* 23(2): 54–61.

Brennan, S. and Winsby, M. (in preparation). “Micro-sanctions: a philosophical exploration.”

Brownstein, M. and Madva. A. (2012). “Ethical automaticity,” *Philosophy of the Social Sciences* 42(1): 68–98. doi: 10.1177/0048393111426402.

Caplan, P. J. (1993). *Lifting a Ton of Feathers: A Woman’s Guide to Surviving the Academic World*. Toronto: University of Toronto Press.

[29] See, for example, Mills (1994).

Correll, J., Park. B., Judd, C. M., and Wittenbrink, B. (2007). "The influence of stereotypes on decisions to shoot." *European Journal of Social Psychology* 37(6): 1102–17. doi: 10.1002/ejsp.450.

Frye, M. (1983). *The Politics of Reality*. Trumansburg, NY: The Crossing Press.

Isaacs, T. (2011). *Moral Responsibility in Collective Contexts*. New York, NY: Oxford University Press.

Kenyon, T. (2014). "False polarization: Debiasing as applied social epistemology" *Synthese* 191: 2529–47.

Kenyon, T. and Beaulac, G. (2013). "Critical thinking and the scope of debiasing in moral thinking," Association for Moral Education, Université du Québec à Montréal, October 26.

Mills, C. (1994). "Do Black men have a moral duty to marry Black women?" *Journal of Social Philosophy* 25(s1): 131–53.

Moore, G. E. (1903). *Principia Ethica.*

Myers, C. K., Bellows, M., Fakhoury, H., Hale, D., Hall, A., and Ofman, K. (2010). "Ladies first?: A field study of discrimination in coffee shops." *Applied Economics* 42(14): 1761–9.

Parfit, D. (1984). *Reasons and Persons*. Oxford: Oxford University Press.

Uhlmann, E. and Cohen, G. (2007). "'I think it, therefore it's true': Effects of self-perceived objectivity on hiring discrimination," *Organizational Behavior and Human Decision Processes* 104: 207–23.

Valian, V. (1999). *Why So Slow? The Advancement of Women*. Cambridge, MA: MIT Press.

Young, S. (2007). Micro Messaging: Why Great Leadership is Beyond Words. New York, NY: McGraw-Hill.

3.4

Discrimination Law, Equality Law, and Implicit Bias

Katya Hosking and Roseanne Russell

1 Introduction

In recent years, governments on both sides of the Atlantic have shown a desire to develop policies informed by the psychology of human behaviour (Thaler and Sunstein, 2009: 14). In this post-financial crisis era, behavioural analyses have contributed to policy reforms in areas such as the composition of corporate boardrooms (Villiers, 2010: 544) and the regulation of financial markets (Avgouleas, 2009). The influence of policies informed by behavioural psychology can also be noted in initiatives to improve citizen welfare such as removing tobacco products from display in large British stores and New York City Mayor Bloomberg's proposed ban on smoking in certain public spaces (a proposal since declared invalid by the State of New York's Supreme Court: *NYC CLASH Inc v NYS Office of Parks, Recreation & Historic Preservation*). Of particular note has been Thaler and Sunstein's 'nudge theory', which claims to offer 'better governance' by incentivizing the electorate to make decisions based on psychological insights about how citizens really behave (2009: 15). Despite criticisms about the theoretical and empirical underpinnings of the nudge movement, government enthusiasm for policy design informed by cognitive psychology shows no sign of waning (Yeung, 2012).

In the area of discrimination, social cognition psychologists have similarly demonstrated how unconscious biases affect our behaviour (Greenwald and Krieger, 2006). Although the science of the implicit bias and behavioural decision-making theorists differ, Kang points to their 'broadly consistent' underpinnings (2005; 1494, n. 21). Both acknowledge the importance of our subconscious in shaping our conscious behaviour. What is striking, however, for the purposes of this chapter, is the starkly different treatment by policy makers of the insights provided by those working in the field of implicit bias. Far from designing new and innovative interventions, policy in Britain is, if anything, drawing back from questions of discrimination and equality. This chapter explores the mechanisms available in British law for working against the effects of implicit bias, and argues that—despite appearances to the contrary—these mechanisms are severely limited by a general underlying approach to law and a specific underlying conception of equality. To the extent that other jurisdictions share this approach to law and this conception of equality, we suggest they are also likely to face these limits.

Britain's discrimination law provisions are contained in the Equality Act 2010 ('EqA'), which consolidates a range of previous legislation prohibiting discrimination on various specific grounds.[1] In Section 2 we analyse the ability of its provisions dealing with direct discrimination, indirect discrimination, and equal pay to provide a remedy for discrimination caused by implicit bias. Since 'sex' represents the characteristic upon which most claims of discrimination are based (for example, in 2011–12, 14,700 sex discrimination cases were disposed of at a tribunal hearing, compared to 4,700 claims based on race and 7,300 based on disability), we illustrate our analysis with examples relating to gender discrimination in employment (Ministry of Justice; 2012, table 2). Although we conclude that nothing in the statutory provisions or case law would prevent a direct discrimination claim being brought on the basis of implicit bias, the practical impossibility of demonstrating, in an individual case, that implicit bias was causally linked to the alleged less favourable treatment renders such a claim almost impossible to prove. While the provisions on indirect discrimination and equal pay offer a different, group-based approach to tackling disadvantage, like direct discrimination, they suffer from procedural limitations.

Section 3 considers whether a newer, proactive model of 'equality law' which applies to the public sector offers greater promise. Moving beyond considerations of proving individual fault, equality law obliges public bodies, in their decision making, to have due regard to the need to eliminate discrimination and advance equality of opportunity. At first glance, equality law offers distinct advantages. It takes the existence of discriminatory inequalities as its starting point and places the onus on the public authority to take action to eliminate these. On closer scrutiny, however, the public sector equality duty also has limitations. This is due,

[1] Section 4 lists the 'protected characteristics' which must not form grounds for discriminatory behaviour, including age, disability, gender reassignment, marriage and civil partnership, pregnancy and maternity, race, religion or belief, sex, and sexual orientation.

we argue, to the 'folk-liberal' orientation of both discrimination law and equality law in Britain, which fails to take adequate account of the structural character of discrimination. In Section 4 we advocate a more substantive vision of equality and suggest two ways by which the current legislative framework might be reformed to achieve this aim.

2 Addressing Implicit Bias within Existing Discrimination Law

2.1 Britain's labour relations context

Britain's provisions against discrimination in employment are located within a broader government labour relations strategy premised on the ideas of flexibility, effectiveness ('enabling employers to manage their staff productively'), and fairness (BIS, 2011: 2). This strategy continues the deregulatory trajectory of the employer–labour relationship over the past thirty years (Davies and Freedland, 1993: 529), which in turn is located within a broader political strategy of neoliberal capitalism (Kotz, 2009: 306). Recent labour law reforms suggest that this trend towards deregulation of business and the employment relationship may be deepening. A 'Red Tape Challenge' has been introduced by the government in a bid to repeal employment regulations perceived as debilitating to business (BIS, 2011: 2), supplemented by measures such as the introduction of fee payments for claimants who wish to bring a complaint in the Employment Tribunal (The Employment Tribunals and Employment Appeal Tribunals Fees Order 2013).

These reforms are rooted in an ideology whose promotion of the values of autonomy and self-sufficiency obscures the power dynamics of the labour relationship. The 'party-devised' contract of employment at the heart of this relationship derives from a nineteenth-century understanding of contract which assumes that autonomous parties negotiate on the basis of roughly equal bargaining power and access to information (Freedland and Kountouris, 2008: 63). This characterization of contract formation is read into employment contracts, even though the contract is invariably drafted by the employer. The myth that each party to the employment contract has negotiated its terms fuels the notion that the employee has freely consented to the terms of engagement (Rönnmar, 2006: 64). Although employment contracts now contain various protections imposed by statute, such as those against discrimination and unfair dismissal, these protections are laid over a contract which fundamentally ignores inequality of bargaining power. This 'blatant preference for liberty over equality' creates a labour environment in which genuine equality and redistribution of resources are subordinated to non-interference with market principles (Fredman, 2002: 25). It is against this backdrop that complaints of discrimination are being made.

Britain's individualistic approach to labour relations becomes clear when the practicalities of bringing a claim are considered. In employment, discrimination law is enforced by individuals who pursue complaints in the Employment Tribunal.[2] It is not possible for claimants to bring class actions, and bodies such as trade unions or Britain's National Human Rights Institution—the Equality and Human Rights Commission (EHRC)—have no power to bring representative actions or to investigate individual cases. Group disadvantage therefore goes unchallenged unless each individual has the means, time, confidence, and courage to contest the alleged discriminatory treatment. A recent study by Busby and McDermont highlighted the particular difficulties faced by individuals seeking to bring claims in the Employment Tribunal without the assistance of trade union representation or a solicitor (2012). Significant hurdles were encountered including the claimants' often fundamental lack of understanding of the Tribunal process and even of the correspondence sent to parties (2012: 175).

This individual approach also makes it much harder to detect the operation of implicit bias which may only be evident in its effects on larger numbers of people. For this reason, the EqA made provision for employees to send a statutory questionnaire to their employer in order to obtain information allowing them to decide whether to commence or continue a claim. This could, for example, include information about the number of women who applied for a post, how many were shortlisted and how many were ultimately successful, thus helping to build up a picture of whether bias may be operating. The force behind this procedure comes, in part, from the fact that a tribunal may draw adverse inferences against an employer who does not respond or whose response is evasive. However, the statutory right in the EqA to ask an employer to respond to questions was abolished on 6 April 2014 (BIS, 2013).

2.2 Direct discrimination and the burden of proof

Direct discrimination is less favourable treatment because of a protected characteristic.[3] In the language of the EqA:

[2] The primary forum for hearing complaints relating to breaches of employment law is the Employment Tribunal. Complex cases, including discrimination cases, are heard by a panel consisting of a legally qualified employment judge and two 'wing members'—one an experienced trade union official or worker representative, and the other an experienced manager or human resources specialist.

[3] Direct discrimination is broader than the US concept of 'disparate treatment' because it includes less favourable treatment which was *caused* by the protected characteristic without being *motivated* by it.

> A person (A) discriminates against another (B) if, because of a protected characteristic, A treats B less favourably than A treats or would treat others. (s13(1))

'Treatment' is not limited to face-to-face interactions but is understood broadly to cover all kinds of behaviour towards someone or dealings with them, and it includes omissions as well as acts (*Alder v Chief Constable of Humberside* [2006] EWCA Civ 1741; EqA, s212). The use of the phrase 'less favourable treatment' instead of 'unfavourable treatment' has come to mean that a claimant has to show that she has been treated less favourably than someone else in relevantly similar circumstances was treated or would have been treated; that is, less favourably than an actual or hypothetical comparator whose circumstances are not materially different (EqA, s23).

Having demonstrated less favourable treatment, a claimant then has to show that the treatment was because of sex. Importantly, treatment is 'because of' a protected characteristic if the protected characteristic is one of the causes of the treatment: it does not matter what the employer's motive was, or even whether the employer knew that the protected characteristic influenced the treatment (EHRC Code, 2011: paras. 3.14 and 3.15). An example of what would constitute direct discrimination is failing to promote a suitably qualified and experienced woman because her employer thinks that she may want to have children and will take time off on maternity leave.

It has long been recognized that unconscious attitudes may be relevant in discrimination cases. As Mr Justice Browne-Wilkinson pointed out in 1981:

> To decide that there has been discrimination in the face of sworn evidence that there was no such discrimination is unpalatable: equally racial discrimination does undoubtedly exist, and it is highly improbable that a person who has discriminated is going to admit the fact, quite possibly even to himself. (*Khanna v Ministry of Defence* [1981] IRLR 331; Employment Appeal Tribunal at 334)

That 'direct discrimination does not have to be, and is usually not, intentional or deliberate' has been confirmed repeatedly (*MOD v Cartner* 2011 EWCA Civ 1516, per Maurice Kay LJ at [11]). As a result, it has also been recognized that it will often be impossible to decide discrimination cases simply on the basis of primary facts established by relatively straightforward evidence; instead, tribunals have to draw inferences from the primary facts:

> ... the legislature is alive to the difficulties of strict proof in this crucially important area and has chosen to meet them by instituting an adjudicative framework in which inference and nuance play a central role. (*Alder v Chief Constable of Humberside* [2006] EWCA Civ 1741 at para. 15, per Sedley LJ)

However, tribunals are restricted in the inferences they may draw, and the ways these restrictions are justified and then worked out in practice reflect the conflicting normative commitments at work in this area of law. The fact that identifying appropriate comparators has become central to these cases complicates matters still further. Although both case law and statute recognize that discrimination often results from unconscious attitudes and assert that victims of such discrimination should be able to obtain a remedy, in practice it is virtually impossible for them to do so.

In civil cases, the burden of proof is ordinarily on the claimant to prove each element of her case on the balance of probabilities.[4] The principal mechanism by which the law takes account of the difficulty of proving discrimination is by shifting the burden of proof: once the claimant has established some initial set of facts the burden is shifted to the respondent to show, again on the balance of probabilities, that the treatment was not because of a protected characteristic. The precise definition of that 'initial set of facts' has developed over time, and for the claimant this definition is of central importance.

In 1991 in *King v Great Britain China Centre*, the Court of Appeal set out guidelines for tribunals on drawing inferences in direct discrimination cases. After emphasizing that 'it is unusual to find direct evidence of racial discrimination', Lord Justice Neill said that

> a finding of discrimination [that is, less favourable treatment] and a finding of a difference in race will often point to the possibility of racial discrimination. In such circumstances the Tribunal will look to the employer for an explanation. If no explanation is then put forward or if the Tribunal considers the explanation to be inadequate or unsatisfactory it will be legitimate for the Tribunal to infer that the discrimination was on racial grounds. ([1991] IRLR 513 at [518])

Although the Court of Appeal held back from describing this as involving a shift in the burden of proof, it did establish that a tribunal would be permitted to infer unlawful discrimination in these circumstances. The guidelines, then, suggested two stages of reasoning: first, was there a difference in race (or sex), and was there discrimination (that is, less favourable treatment on grounds of race or sex)? Second, has the employer provided an adequate and non-discriminatory explanation for the treatment? At the first stage, only the *possibility* of race or sex discrimination is established. The presence or absence of a satisfactory explanation at the second stage settles whether there actually was unlawful discrimination.

The tribunal in *King v Great Britain China Centre* had found it relatively straightforward to conclude that the claimant, who was from a Chinese ethnic

[4] The phrase 'balance of probabilities' refers to the standard of proof in civil cases, which is a lower standard of proof than the criminal 'beyond reasonable doubt' standard. In civil cases, the claimant only has to show that it is more likely than not that her assertions are true.

background, had been treated less favourably when she was not shortlisted for a position because it was able to identify actual white comparators among the people who had been shortlisted. The claimant's qualifications and experience were not significantly different from those of the shortlisted candidates, and the tribunal held that the selection criteria had not been applied consistently. In addition, none of the other four Chinese British applicants had been shortlisted, nobody from a Chinese ethnic background had ever been employed by the Centre, and the tribunal noted that that 'the evidence of the current director on this point was considered . . . to be unsatisfactory' (at [516]).

In 2001, the law in relation to burden of proof was altered in response to changes in EU law. In order to make implementation of the equal treatment principle more effective, the Burden of Proof Directive required EU member states to

> take such measures as are necessary . . . to ensure that, when persons who consider themselves wronged because the principle of equal treatment has not been applied to them establish . . . facts from which it may be presumed that there has been direct or indirect discrimination, it shall be for the respondent to prove that there has been no breach of the principle of equal treatment. (Article 4(1), Council Directive 97/80/EC of 15 December 1997)

According to the guidelines in *King v Great Britain China Centre*, a tribunal was merely *permitted* to shift the burden of proof, but the position is now that the tribunal *must* shift the burden of proof once the claimant has proved the relevant facts. The guidelines were duly amended and expanded in *Igen v Wong* ([2005] EWCA Civ 142) and approved in that form by the Supreme Court in *Hewage v Grampian Health Board* ([2012] UKSC 37).

It has not always been straightforward to identify what facts are sufficient to shift the burden of proof, however. It is not enough for the claimant simply to show a difference in treatment and a difference in protected characteristic:

> This does not mean . . . that in a case involving alleged race discrimination, it will be sufficient at this first stage for an applicant who is black simply to show, for example, that a white comparator was promoted to a post for which he had applied. In view of the . . . need for the relevant circumstances in the applicant's case to be the same or not materially different in the case of the comparator, the applicant in such a case would have to show not only that he met the stated qualifications for promotion to the post, but that he was at least as well qualified as the successful candidate. (*Dresdner Kleinwort Wasserstein v Adebayo* [2005] IRLR 514)

The cases suggest that claimants will only succeed in cases where some further set of facts, beyond a difference in treatment and a difference in protected characteristic, provides an evidential base for inferring that bias has influenced a particular decision. For instance, in *EB v BA*, that additional evidence was the difference in treatment before and after EB's transition to her preferred gender ([2006] EWCA Civ 132). In *Anya v University of Oxford*, it was a history of hostile treatment from a panel member together with a failure to apply the employer's own equal opportunities policy to the recruitment process ([2001] EWCA Civ 405). In these cases, then, the claimant and respondent already had a history in which patterns of behaviour could be seen.

However, implicit bias will sometimes—even commonly—operate where there is no such history. In that context, a tribunal will have few or no further facts or surrounding context with which to support its inferences. Barrister Naomi Cunningham imagines just such a 'pure' case in her hypothetical recruitment interview for a vacant mathematics chair (2006: 280–1). In her scenario, the equally well qualified male and female candidates are both capable of fulfilling the requirements of the role, and both impress the panel at interview. The all-male selection panel chooses, after much debate, to appoint the man. On what basis could a tribunal find that implicit bias has caused the woman not to be appointed? The chair of the selection panel is honest, and agrees at a subsequent tribunal hearing that the decision was so finely balanced that another panel might have arrived at a different decision. In this context, Cunningham argues that

> [t]here is no evidence here that could satisfy a reasonable tribunal on the balance of probabilities that the complainant has suffered discrimination. There is nothing to suggest the existence of a prejudice. The witness is honest and straightforward: he does not attempt to deny the undeniable facts of the society of which he is part, but neither does he show any sign of a positive desire—conscious or otherwise—not to appoint a woman. If the burden is on the claimant she must lose. (2006: 281)

Yet if the burden shifts to the employer to show that its decision was not motivated by implicit bias,

> . . . it is little better placed. The panel members may satisfy the tribunal that they had no conscious intention to discriminate, but on what evidence could they hope to prove an absence of unconscious discrimination? (2006: 281)

Thus the outcome of this case turns on the burden of proof and, according to Cunningham, the burden will shift in this case, and the result will be unfair to the employer. However, this interpretation has not been borne out by the case law. On the contrary, the scenario as described would not be sufficient to shift the burden to the respondent. Cunningham appears to assume that the appropriate comparator would be the successful candidate, and of course it is clear that the claimant has been treated less favourably than he has. But it is not enough that the claimant is female and the successful candidate male: something more is required. As Lord Justice Sedley put it in *Anya*:

> The choice between these two comparably well qualified candidates depended entirely on how the panel viewed their personal and professional qualities. Such a judgment is notoriously capable of being influenced, often not consciously, by idiosyncratic factors, especially where proper equal opportunity procedures have not been followed. *If these are to any significant extent racial factors, it will in general be only from the surrounding circumstances and the previous history, not from the act of discrimination itself, that they will emerge.* ([2001], EWCA Civ 405; [2001], IRLR 377, at 382–3; (emphasis added)).

Yet here, by hypothesis, no previous history is presented. If the panel has judged the man as the more meritorious candidate, however narrowly, and the tribunal found the panel chair to be a credible witness, it will have no basis for shifting the burden because it will not have found 'facts from which the court could decide, in the absence of any other explanation, that [the panel] contravened the provision concerned' (EqA 2010, s136(1)). As Mummery LJ opined in *Madarassy v Nomura International plc* ([2007] EWCA Civ 33 at [56]):

> The bare facts of a difference in status and a difference in treatment only indicate a possibility of discrimination. They are not, without more, sufficient material from which a tribunal 'could conclude' that, on the balance of probabilities, the respondent had committed an act of unlawful discrimination.

Whether the burden shifts in Cunningham's hypothetical case tells us which errors the litigation process is more willing to tolerate. Do we leave the burden with the claimant and risk her not having a remedy, or are we more willing to find that someone has discriminated, even unconsciously, when they have not? It turns out that in 'pure' implicit bias cases the burden is left with the claimant.

2.3 Indirect discrimination

Protection against a second form of discrimination may offer more potential to address the effects of implicit bias. 'Indirect' discrimination arises where a provision, criterion, or practice ('PCP') appears neutral but, in practice, would put people who share a protected characteristic at a particular disadvantage (EqA s19).[5] For instance, it captures what is problematic about a requirement that all employees must work from 9 a.m. to 5 p.m.: the requirement is neutral on its face, but it puts people with caring responsibilities, the majority of whom are women, at a disadvantage. However—unlike direct discrimination, which cannot be justified except in cases of age—this potentially discriminatory treatment will

[5] This most closely resembles the US concept of 'disparate impact'.

not be unlawful if the employer can show that applying the PCP was 'a proportionate means of achieving a legitimate aim.' For example, it may be a legitimate aim for a business to maintain adequate workforce levels during operating hours. Whether it is proportionate to meet that aim by insisting that every employee works full time 9 a.m. to 5 p.m. will depend on factors such as the nature and size of the business.

Indirect discrimination therefore tackles practices which have problematic effects even when the practices are not obviously discriminatory and the effects are unintended, which is one of the patterns we might expect to see when implicit bias is operating. Once again, however, there are difficulties with this model when it comes to providing a remedy for the operation of implicit bias in an individual case.

To bring a claim of indirect discrimination a claimant must identify a neutral PCP which puts (or would put) women *as a group* at a disadvantage and which also puts (or would put) her *as an individual* at a disadvantage. Identifying and formulating precisely the correct PCP is vital and can be surprisingly tricky. But a PCP like 'must work 9–5' does not create its discriminatory effect through the operation of implicit bias—its content and application are explicit. It is just that women are, as a matter of current social fact, less likely to be able to meet the requirement.

On the other hand, sometimes there will be nothing in the content of the PCP which ought to create disadvantage. In the case of a loose recruitment practice where explicit selection criteria have not been used or inconsistent questions have been asked of candidates, the practice is not intrinsically disadvantageous. The disadvantage might only become clear over time when the numbers reveal that the practice is resulting in a smaller proportion of women being selected. Although evidence of uneven recruitment might *suggest* the operation of implicit bias, it is almost impossible to show. In a case of indirect sex discrimination, the claimant must convince the tribunal that a PCP exists, that it disadvantaged women generally and that she was or would be disadvantaged. Only after all three factors have been demonstrated will the burden of proof shift to the employer 'to provide both explanation and justification' (per Langstaff J in *Dziedziak v Future Electronics Limited* UKEAT/0270/11 at [42]). Moreover, while the non-appointment of a woman in the context of a pattern of similar behaviour is redolent of group disadvantage, such a case would most likely be pursued as a direct discrimination claim; that is, the claimant would simply assert that she has been treated less favourably because of sex. The statistical evidence thus becomes an important background fact to help shift the burden of proof, but any remedy is limited to the individual claimant.

2.4 *Equal pay*

A third form of discrimination in which implicit bias might operate against women in the employment context is concerned specifically with discrimination in how pay or other contractual terms are awarded: 'equal pay'. It has been recognized that 'there may well be a basis for inferring that the employer has (in the past) had a subconscious attitude that women do not need to earn as much as men and that the present pay arrangements are a legacy of that attitude' (per Smith LJ in *Gibson v Sheffield City Council* [2010] ICR 708 at [68]). The equality of terms provisions in the EqA, largely repeated from the Equal Pay Act 1970, appear 'deceptively simple' (Steele, 2010: 264). Where men and women perform 'like work'—work that has been rated as equivalent under an analytical job evaluation scheme, or work of equal value—it is presumed that they should be paid the same (EqA s65(1)(a)–(c), s66). To achieve this, an 'equality clause' is implied into the contract of employment whereby a woman's less favourable terms are modified to those of the comparable man (EqA, s66). There may, however, be a genuine and reasonable explanation[6] for why a woman is being paid less than a male co-worker. If the difference is due to *direct* discrimination on grounds of gender, it cannot be justified and the employee will win her claim. If, however, the claim is based on *indirect* discrimination—where gender is not the *cause* of the treatment but the seemingly neutral pay practice has a disparate *effect* on one sex—the employer might still defend its practice. To do this, it will have to show that the practice is objectively justified, that it is a proportionate means of achieving a legitimate aim.

The question of when employers will be required to justify any difference in pay demonstrates how even the apparently more progressive model of equal pay is weak when dealing with the effects of implicit bias. Imagine that, statistically, employees in a mainly (or indeed exclusively) female group are paid less than employees belonging to a mainly male group (*Enderby v Frenchay Health Authority* [1994] ICR 112). It is plausible that implicit bias could have informed the differences in pay. Requiring an objective justification for the pay differential where there is a statistical gender imbalance shifts the burden of disproving discrimination firmly to the employer. Yet the case of *Armstrong v Newcastle upon Tyne NHS Hospital Trust* ([2006] IRLR 124) suggests that even clear evidence showing that groups of women earn less than comparable groups of men may be insufficient to shift the burden of proof. There it was held that even

[6] A 'material factor' other than sex: EqA, s69.

where the disadvantaged employees share the same gender it should not be assumed that the difference in pay is due to sex if the statistical imbalance could be explained by other reasons. In one case where care workers (who were predominantly women) were paid less than an almost exclusively male group of street cleaners and gardeners, despite these roles having been rated as equivalent, the difference was held not to be due to sex. This was because care work was not considered to be susceptible to productivity incentives in the same way as gardening or street cleaning, and so bonuses were not available to care workers. The Court of Appeal has since overturned this decision (*Gibson v Sheffield City Council*) and held that where the disadvantaged group was dominated by one gender and the advantaged group dominated by the other, a 'sex taint' would *likely* be presumed and the factors influencing the pay differential would need to be objectively justified. It did not, however, go so far as to say that objective justification would always be required in such circumstances, and this issue remains unclear under the EqA.

2.5 Evaluating existing discrimination law

Current discrimination law offers a few possible openings for addressing disadvantage caused by implicit bias. First, the positive action provisions in ss158 and 159 of the EqA allow employers to provide schemes such as additional training or mentoring for members of underrepresented or disadvantaged groups. Since 6 April 2011, a protected characteristic can also be used as the 'tipping factor' in recruitment but only when both candidates are 'as qualified' as each other. These positive action provisions have the potential to address some kinds of inequality which might arise from the operation of implicit bias. For instance, if stereotype threat contributes to the fact that women in academia tend to apply for promotion later than men do, then mentoring from senior female academics might help to reduce the strength of the stereotypes in question. Similarly, in Cunningham's mathematics scenario, the chair of the selection panel admitted the decision was so finely balanced that another panel might have taken a different view: the selection panel was in precisely the position where it might now be permissible to take gender into account to tip the balance, given the underrepresentation of women at senior levels in mathematics departments.

However, these provisions are voluntary. They *permit* an employer to take such steps without acting unlawfully but they do not *require* employers to do so. They are also limited by the requirement that an employer must have evidence that members of the relevant group are disadvantaged or underrepresented before taking positive action. Such evidence might be relatively straightforward to acquire in some contexts—for instance, the small number of male primary (elementary) schoolteachers—but it creates an additional hurdle for risk-averse employers. The likely impact of the 'tipping point' provision is also limited by the fact that a characteristic such as gender may only be given preference once a selection panel has decided that two candidates are as qualified as each other; but implicit bias may affect the selection panel members so that (for instance) women are not judged to be as qualified as men. Thus the apparently more substantive approach to equality suggested by the positive action provisions is severely curtailed by their voluntary nature and the prerequisite that both candidates are considered to be 'as qualified' before a characteristic such as gender is given preference.

The second way in which implicit bias might be addressed is through the power of Employment Tribunals, having concluded that there has been discrimination in an individual case, to make recommendations to the employer about measures to counter the negative effects of discrimination (EqA 2010, s124(3)). This is potentially far-reaching. For example, it can include requiring an employer to audit its employment practices and carry out equality and diversity training (*Lycée Français Charles de Gaulle v Delambre* UKEAT/0563/10). If the employer fails, without reasonable excuse, to comply with the recommendation, additional compensation may be awarded to the claimant (EqA 2010, s124(7)). This allows the Tribunal to consider whether the circumstances of an individual's case suggest that her claim reflects a systematic problem within that institution, and to recommend steps which it hopes will prevent the same discrimination happening again.[7] This power to make 'wider' recommendations has been repealed with effect from 1 October 2015 and now recommendations may only relate to the individual claimant (Deregulation Act 2015, s2).

The third innovation is the government's introduction of equal-pay audits on 1 October 2014 (The Equality Act (Equal Pay Audits) Regulations 2014). Where an employee has won a claim for equal pay, the employer will be required to carry out a more general audit of employee pay to see whether unequal pay was an aberration or whether there are more general, deep-rooted pay disparities between male and female employees. Arguably, however, these reforms do not go far enough given that they will only be used where a claimant wins her case and tribunals have no authority to order an audit in certain circumstances including where the disadvantages of an audit would outweigh its benefits

[7] Indeed, people who bring discrimination cases will sometimes say explicitly that they are doing it partly so that nobody else has to go through the same thing. In that sense, when a tribunal makes wider recommendations it may also be satisfying one of the individual claimant's desires.

(Regulation 3). The government's commitment to keeping proactive reporting purely voluntary is clear from its 'Think, Act, Report' initiative (GEO, 2011).

The tools offered by discrimination law to counter disadvantage arising from the operation of implicit bias are therefore limited, and some have suggested that this is partly due to the model of equality currently recognized by British law. As Nicola Lacey suggests, in the context of discrimination law a goal of 'equality of opportunity' functions more as a piece of 'political rhetoric' rather than a well worked-out theory (1987: 414). What you think is necessary to create equality of opportunity—whether you understand it in a procedural or substantive sense—will depend on how far you recognize the extent of disadvantage and subordination experienced as a result of group membership. Our legal and legislative system, however, is framed by a fundamental belief in people's ability to make free and conscious choices. It therefore requires weighty evidence to overcome the presumption that inequality of outcomes is the result of individual choice. That individual legal subjects are deemed of equal value in the eyes of the law entitles them to consistent treatment with similarly placed others (Hepple, 2008: 7), but the formal, procedural model of equality which results from a strict adherence to the principle of equal treatment is 'incapable of removing the apparently neutral barriers . . . which put ethnic minorities and women at a disadvantage' (Hepple, 2008: 4).

Sandra Fredman suggests that the pursuit of substantive equality, by contrast, involves

> four overlapping aims . . . First . . . to break the cycle of disadvantage associated with status or out-groups . . . Secondly . . . to promote respect for dignity and worth, thereby redressing stigma, stereotyping, humiliation and violence because of membership of an identity group . . . Thirdly, it should not exact conformity as a price of equality . . . [but] aim to achieve structural change . . . Finally, substantive equality should facilitate full participation in society, both socially and politically. (2011a: 25)

Implicit bias is implicated directly in maintaining the first, second, and fourth forms of inequality, and also indirectly in the third. As we have seen, existing discrimination law, at best, may help to reduce one of the mechanisms which perpetuates the 'cycle of disadvantage'—direct discrimination—but it is least likely to be effective in just the circumstances where implicit bias may have played a role in that discrimination. It fails to deliver the redistribution necessitated by a commitment to substantive equality because distributive outcomes are deemed to reflect a person's choices, yet it is precisely because of the operation of implicit bias that substantive equality is needed. In the last fifteen years, a new form of legal regulation has been introduced which explicitly aims at promoting (some form of) substantive equality. We will call this 'equality law'.

3 Can Equality Law Counter Implicit Bias More Effectively?

3.1 The public sector equality duties

The public sector equality duties created a wholly new form of regulation for equality in the UK by requiring public bodies to have 'due regard' to the need to promote equality. They are often characterized as positive duties because they involve taking deliberate steps and not merely refraining from unlawful discrimination. A duty in relation to 'race' was introduced in Britain in the Race Relations (Amendment) Act 2000 (s71) after the Lawrence Inquiry[8] found that the Metropolitan Police Service was 'institutionally racist'. Institutional racism was defined as

> [t]he collective failure of an organization to provide an appropriate and professional service to people because of their colour, culture, or ethnic origin. It can be seen or detected in processes, attitudes and behaviour which amount to discrimination through unwitting prejudice, ignorance, thoughtlessness and racist stereotyping which disadvantage minority ethnic people. (MacPherson, 1999: para. 6.34)

This definition therefore includes some of the potential consequences of implicit bias. Further public sector duties were introduced in the Disability Discrimination Act 2005 (s49) and for gender in 2006 (amending the Sex Discrimination Act 1975 to include s76A). The EqA 2010 drew these together into a single Public Sector Equality Duty covering all the protected characteristics except marriage and civil partnership. Under s149, public authorities must, in exercising their functions, have due regard to the need to eliminate discrimination, advance equality of opportunity and foster good relations between people who share a protected characteristic and people who do not share it. In relation to equality of opportunity, it explains further:

> (3) Having due regard to the need to advance equality of opportunity . . . involves having due regard, in particular, to the need to
>
> (a) remove or minimize disadvantages suffered by persons who share a relevant protected characteristic that are connected to that characteristic;
> (b) take steps to meet the needs of persons who share a relevant protected characteristic that are different from the needs of persons who do not share it;

[8] Stephen Lawrence, a black British man, was murdered in 1993 in a racist attack. The police investigation of the murder was blighted by serious and repeated failings, and the MacPherson inquiry into the police's handling of the investigation revealed evidence of institutional racism. In 2012 two white men who were part of the original group of suspects arrested in 1993 were finally convicted of murder.

(c) encourage persons who share a relevant protected characteristic to participate in public life or in any other activity in which participation by such persons is disproportionately low.

Rather than relying on enforcement by means of individuals bringing complaints, then, the public sector duties aim to promote change by improving the range and quality of data about the experiences of members of particular groups and by requiring systematic consideration of equality matters as part of the decision-making processes of public bodies.[9] For that reason, they should be able to address disadvantage and inequality which arises from the operation of implicit bias but which is impossible to challenge in any individual case.

In order to implement the general duty effectively, specific equality duties are imposed on a more restricted list of public bodies setting out procedures the organization is expected to follow as a means of meeting its obligations under the general duty. These have included gathering, analysing and publishing information about the impact of their policies and practices on different groups of people—often referred to as conducting 'equality impact assessments'—consulting with members of the relevant groups, and publishing an equality scheme setting out their objectives and action plans in relation to meeting the general duty.

3.2 Enforcement

Only the EHRC is able to enforce the specific equality duties; however, it has not pursued a particularly active enforcement policy. Where it has, for instance, issued compliance notices it is difficult to find out what happened subsequently.[10]

Individuals can seek to enforce the general equality duty by means of judicial review. Judicial review in the UK is the mechanism by which people and organizations can challenge the decisions of public bodies but, generally speaking, the circumstances in which such a challenge is likely to be successful are very limited and are almost always procedural.[11] The challenger will usually have to show that the decision is '. . . so outrageous in its defiance of logic or of accepted

[9] This is often referred to as 'embedding' or 'mainstreaming' equality.

[10] An example of enforcement can be seen at <http://www.equalityhumanrights.com/cy/commission-issues-gender-equality-duty-compliance-notices-three-local-authorities> (accessed 25 September 2015), but there are no updates at. <www.equalityhumanrights.com/legal-and-policy/enforcement>. Instead, <www.equalityhumanrights.com/legal-and-policy/enforcement/legal-enforcement-case-studies/> (accessed 25 September 2015) offers a 'flavour' of the EHRC's enforcement activities.

[11] It is a much more limited power than that established in the USA under *Marbury v Madison* 5 U.S. 137, where courts review the constitutionality of legislative action and have the power to declare it void.

moral standards that no sensible person who had applied his mind to the question to be decided could have arrived at it' (*Re Council of Civil Service Unions* ('the GCHQ case') [1985] AC 374, per Lord Diplock at [410]). In other words, all public decision makers are required to do is avoid blatant irrationality and outrageous immorality.

However, in 2006 *R (Elias) v Secretary of State for Defence* established that obligations under the general equality duty could also be enforced by means of judicial review, and it was therefore confirmed that the equality duties were held to impose an additional requirement on decision making by public bodies ([2006] EWCA Civ 1293). In practice, this additional requirement will be met if the specific duties are fulfilled. For instance, in a case challenging Ealing Council's decision to withdraw funding from Southall Black Sisters, a group which supported minority ethnic women experiencing domestic violence, the court emphasized the requirement to conduct a meaningful equality impact assessment at an early stage in the decision-making process (*R(Kaur and Shah) v London Borough of Ealing* [2008] EWHC 2062 (Admin)). The court held that—even though it was not blatantly irrational or immoral—the council's decision could not stand because it had been made without giving proper consideration to the evidence that minority ethnic women were more likely to be the victims of domestic violence. This shows that the courts are now willing to hold public bodies to a higher standard of decision making where there may be a significant negative impact on equality.

3.3 *Equality law and implicit bias*

Positive equality duties, then, seem to avoid many of the problems with discrimination law. According to Mark Bell, what the Lawrence Inquiry meant to convey in its definition of institutional racism was that

> aside from any evidence of overt racism on the part of the officers involved, the overall organizational response did not treat his murder in the manner that would have been adopted for a white victim. Consciously or unconsciously, stereotypes were applied and implicit assumptions made which negatively affected the conduct of the investigation. (2010: 674)

So the equality duties ought precisely to address the phenomenon of implicit bias. Enforcement is not restricted to claims brought by individuals and there is no need for a finding that someone is at fault to trigger the obligation to act. They are aimed at preventing discrimination, not merely at responding once it has taken place, and in addition they aim to remove disadvantage. Mechanisms such as equality impact assessments are capable of capturing the effects of implicit bias across a population, in that they require an analysis of monitoring data in order to identify disproportionate impact of policies on particular groups. Unlike in an individual indirect discrimination case, there is no requirement to identify a 'provision, criterion, or practice' which creates the disadvantage before considering whether action can be taken to remove or reduce the disadvantage.

In addition, because they do not involve finding fault, there may be less reluctance to acknowledge an organization's failings and so these duties have the potential to increase transparency about an organization's practices and plans. The hope is that awareness of the possibility of public scrutiny encourages decision makers to be more scrupulous about the fairness of their procedures and the information they rely on. The equality duty might therefore be seen as a form of reflexive regulation in which the law is used as a vehicle to trigger the system to modify itself rather than prescribing fixed rules and sanctions (Hepple, 2011; McCrudden, 2007: 259).

In practice, however, the potential of equality duties is limited in several ways. First, the duty only applies to listed public bodies and not to the private sector unless exercising a public function, which has been interpreted narrowly (Fredman, 2011b: 415). In addition, although the employment functions of a public body are covered by the equality duties, employment decisions cannot be challenged by means of judicial review: in their capacity as employers public bodies are acting under the contractual employment relationship and not relying on their public law powers. This may mean that the equality duties are even less likely to be effective in the employment context. Secondly, although the equality duties avoid the question of whether anyone is at fault, concerns about reputation management might make organizations reluctant to publish equality impact assessments, limited in the numbers they do publish, and cautious about the problems they have identified.

Thirdly, and more fundamentally, the character of the duty itself is also weak. In particular, it is formulated in a way which shifts the emphasis back from outcomes to processes: instead of imposing a duty to take steps to advance equality, it requires only that organizations have due regard to the need to advance equality. This means that ultimately organizations can leave disadvantage in place if they can offer justification. It also limits the remedies which are available, because even if a court finds that an organization failed to have due regard in making a decision or in applying a policy, and sets it aside, the decision is then returned to the organization to reconsider. Although the organization may reach a different conclusion once it takes equality into account, it is not required to do so: it may simply reach the same conclusion but with a more detailed account of its reasons.

4 Suggestions for Reform

The apparently progressive legal frameworks of British discrimination law and equality law ought, on the face of it, to be capable of tackling implicit bias. Their implementation, however, is significantly and systematically constrained by an underlying conception of discrimination as an individual harm. Efforts to remove these constraints are hampered by the terms of legislative debate, structured by a vague 'I-know-it-when-I-see-it' liberalism, which restricts the kinds of arguments and reasons which can be offered for legislating. There is a strong presumption in favour of a (somewhat muddled) neutrality of the state, together with a (similarly muddled) division between public and private spheres. In the private sector this is particularly problematic when it comes to regulating for equality, as companies are considered themselves to be private legal persons with rights and responsibilities. As Fineman notes:

> Markets are constructed as public (and therefore under a different, competitive set of norms) when contrasted with the family, but as private (and therefore not easily susceptible to public regulation) when paired with the state. The market reaps the best of both spheres. (Fineman, 2005: 22)

This presents particular problems for challenging implicit bias. Countering such subconscious prejudice is, argue Hardin and Banaji, 'especially insidious in a society that celebrates, evaluates, and is organized around individual meritocracy' (2013: 23), yet this is precisely the context in which the British debate is framed. This supports an individualistic and fault-based approach to discrimination but downplays the extent to which systematic and structural disadvantage can undermine autonomy. As Moreau argues: '. . . the interest that is injured by discrimination is our interest in . . . deliberative freedoms . . . freedoms to have our decisions about how to live insulated from the effects of normatively extraneous features of us, such as our skin color or gender' (2010: 147). Discrimination which entrenches historic patterns of disadvantage and subordination leads, in Moreau's view, to a restriction in the exercise of those supposedly valued liberal principles of freedom and autonomy. It is precisely because discrimination is '. . . a social problem that entrenches the subordinated and disadvantaged status of particular, socially salient groups . . .' (Bagenstos, 2007: 490) that legal remedies informed by narrow conceptions of liberalism and formal models of equality are ill equipped to deal with implicit bias. Instead, we need to move beyond formal models of equality which, at their most reductive, simply call for 'sameness' of treatment (Squires, 2005: 369) to more substantive models of equality which attempt to challenge patterns of subordination and allow for the genuine exercise of deliberative freedoms. This idea is controversial because it depends, in part, on the state assuming a more active role in distributing societal goods and resources in an environment in which the autonomous legal subject is lauded. Disagreement also persists regarding how such resources should be allocated.

The hurdle of showing facts from which discrimination may be inferred before the burden shifts to the employer to disprove a direct discrimination claim also reinforces the requirement that someone must be at fault. It demands that the claim be seen from

> the 'perpetrator perspective'—the notion that what matters is whether the person accused of discrimination was at 'fault', and not whether the person accused of discrimination actually caused or contributed to group-patterned harm. (Bagenstos, 2007: 488, drawing on Alan Freedman)

To counter the group harm of implicit bias it is the dimensions of substantive equality identified by Freedman—redistribution, recognition, structural transformation, and participation—that are needed. The positive equality duties should offer a real opportunity in this regard, but in practice they have been trimmed and constrained to fit back with the liberal framework and their effects therefore remain limited.

In order to begin tackling the effects of implicit bias the law needs to move beyond seeing discrimination simply as an individual harm. Following Lawrence, viewing implicit bias as something more like a public health concern would liberate us from having to locate a blameworthy moral agent (1987: 321). As Shin comments:

> If this means changing our understanding of discrimination from an agent-centered, moralistic conception to a predominantly psychosocial, diagnostic one, then perhaps our wisest response might be to bite the necessary bullets and avow that latter conception. If unconscious discrimination really is best characterized as akin to passing on an infectious disease, then maybe the law should approach the problem of such discrimination not in the traditional manner of assigning individual responsibility and blame, but much more in the manner of addressing an issue of public health. (2010: 101)

Such a reformulation of the accepted nature of discrimination requires, in our view, at least two practical changes to the structures by which claims of discrimination are made. The first amends the definition of direction discrimination from the current formulation of *less* favourable treatment because of a protected characteristic to *un*favourable treatment because of a protected characteristic. This subtle change is borrowed from the current British definition of pregnancy-related discrimination, which requires proof of *un*favourable treatment only and so avoids the need for a comparator (EqA, s18(2)). As we have already seen, a direct discrimination claimant ordinarily has to show *both* that she has experienced less favourable treatment than a strictly defined comparator, *and* that this treatment was because of sex. In contrast, someone alleging pregnancy-related discrimination only needs to show that she has experienced unfavourable treatment—treatment which put her at a disadvantage—and that this treatment was because of her pregnancy. Claimants might still wish to show how a comparable non-pregnant person has been treated, as one evidential route to showing that the treatment was because of her pregnancy, but there is no legal requirement to do so in order to succeed in a claim.

The consequences of this definition might best be illustrated by example. In a promotion exercise where candidates are scored based partly on their client lists, a woman returning from maternity leave is scored lower than colleagues because not all clients have yet been returned to her. It may be difficult to show that she has been *less* favourably treated *than others* have been or would have been, because it is not clear what circumstances we should pick out as being comparable to returning from maternity leave. However, it is easier to show that this is *un*favourable treatment which is because of her maternity leave.

To see the implications for cases involving implicit bias, suppose that two university lecturers have administrative portfolios which their institution's workload matrix rates as equivalent. The man's tasks tend to focus on research administration and postgraduate students, the woman's on personal tutoring and pastoral care. Over the years, heads of department have allocated administrative tasks in this way based on who they regarded as most suited, and these allocations were affected by implicit gender bias. The man develops a reputation in the department as a serious scholar who prioritizes research, the woman as a collegiate team-player who is well liked by students. The man is promoted and the woman is not. Given the requirement for a comparator, the employer would argue that the woman has not been discriminated against: she has not been treated less favourably than a man with the same administrative profile as her would have been treated. But the problem is not that a man with her profile would have been promoted; the problem is that no man would have ended up with her profile. It is precisely because of the operation of implicit bias that the woman and the man are no longer suitable comparators.

Altering the definition of direct discrimination to *un*favourable treatment would allow the courts to move beyond the current unsatisfactory state of affairs where a case can turn on the construction of the correct actual or hypothetical comparator. Instead, the courts could concentrate on the cause of the unfavourable treatment. In our academic promotion case, removing the requirement for a comparator means that the story the woman needs to tell is a different one.

Rather than trying to show how a hypothetical man would have been treated, she just has to demonstrate that the allocation of less prestigious administrative tasks, which affected her reputation and her chances of promotion, was unfavourable and was as a result of gender. This should be easier to demonstrate by, for instance, looking at the patterns of workload allocation in the department. While this may involve a comparative exercise, this is not the same as the law's current insistence on the woman having to first persuade the courts that an actual or hypothetical comparator has been found or constructed against whom she can compare her treatment.

Our second proposal seeks to move beyond attributing blame to acknowledging the societal influences which inform our implicit biases. In a case of 'pure' implicit bias—such as that described by Cunningham—we suggest that the burden of proof should shift to the employer simply on the basis of the prima facie case, without requiring the 'something more' referred to by Lord Justice Mummery in *Madarassy*. Recall Cunningham's concern, in that context, that

> [t]he panel members may satisfy the tribunal that they had no conscious intention to discriminate, but on what evidence could they hope to prove an absence of unconscious discrimination? (2006: 281)

We suggest that the employer should precisely be expected to show that it took all the steps that it was reasonable for it to take to minimize the operation of implicit bias.[12] This test does not rely on a claim that the employer is to blame for the operation of implicit bias in society but suggests that the employer is nonetheless responsible for minimizing its impact once it is aware that it may be operating. The test would take seriously the idea that an employer has a responsibility to its employees to take reasonable steps to protect them from the harmful effects of discrimination. This echoes Fredman's view of the public sector duty that 'even though not responsible for creating the problem in the first place, such duty bearers become responsible for participating in its eradication' (2001: 164). The proposed test is not a wholly novel one; for instance, in cases where a claimant is seeking to hold an employer vicariously liable for the actions of a colleague, the employer has a defence where it can show that it 'took all reasonable steps to prevent' the colleague from doing the alleged act or even 'from doing anything of that description' (EqA, s109(4)). The nature of the steps that should be taken may be altered depending on the size and resources of the employer, with less being expected of employers with more limited resources such as small employers with

[12] Similarly, Green suggests developing a structural account of disparate treatment which 'holds employers directly liable for organizational structures and institutional practices that unreasonably enable the operation of discriminatory bias in the workplace' (2003: 93).

no in-house employment or anti-discrimination expertise. For larger employers, steps might include addressing structural barriers such as redacting names from application forms at short-listing stage, conducting training for all employees on the nature and operation of implicit bias, and a commitment to taking lawful positive action in appropriate cases. If employers can demonstrate to the satisfaction of the courts that they have taken all reasonable steps to minimize the operation of implicit bias, the claim would not succeed.

5 Conclusion

Nothing in the terms of British discrimination or equality law prohibits grounding a claim of discrimination in the operation of implicit bias. Yet, due to the operation of structural barriers external to the legislative content, discrimination as a result of implicit bias is almost impossible to address. The deeply individualistic interpretation of discrimination coupled with a formal model of equality has resulted in a remedial framework centred on individual remedy, ascribing blame, and ensuring little more than equivalent treatment. The commitment to deregulation within a neoliberal economic framework intensifies this approach. As a result, although we know that discrimination often operates to entrench existing group subordination, the current legislative framework is ill equipped to counter this. Our suggestions for reform—framing direct discrimination as unfavourable treatment and obliging employers to show that they took all reasonable steps to minimize the operation of implicit bias—are intended to shift the law of direct discrimination from its current bias towards the standpoint of the alleged perpetrator to a position informed by behavioural realism. The benefit of such reforms is that they are relatively easy to achieve through legislative amendment which is not reliant on wholesale structural displacement of current social and economic systems, although persuasive arguments might well be advanced for more radical changes. The receptiveness of the current government to policy reform grounded in cognitive psychology offers a significant opportunity to call for legislative reform to our discrimination and equality laws.

References

Avgouleas, E. (2009). 'The global financial crisis, behavioural finance and financial regulation: In search of a new orthodoxy'. *Journal of Corporate Law Studies* 9: 23–59.

Bagenstos, S. R. (2007). 'Implicit bias, "science", and antidiscrimination law'. *Harvard Law and Policy Review* 1: 477–93.

Bell, M. (2010). 'Judicial enforcement of the duties on public authorities to promote equality'. *Public Law* 672–87.

Busby, N. and McDermont, M. (2012). Workers, marginalised voices and the employment tribunal system. *Industrial Law Journal* 41: 166–83.

Cunningham, N. (2006). Discrimination through the looking-glass: Judicial guidelines on the burden of proof. *Industrial Law Journal* 35: 279–88.

Davies, P. L. and Freedland, M. R. (1993). *Labour Legislation and Public Policy: A Contemporary History.* London: Clarendon Law Series.

Department for Business, Innovation and Skills (BIS) (2011). 'Flexible, effective, fair: promoting economic growth through a strong and effective labour market'. BIS, London, October.

Department for Business, Innovation and Skills (BIS) (2013). 'Enterprise and Regulatory Reform Act 2013. Commencement Provisions within the Act: Indicative Timetable'. BIS, London, June.

Equality and Human Rights Commission (EHRC) (2011). 'Equality Act 2010 Statutory Code of Practice: Employment'. The Stationery Office, London, January.

Fineman, M. A. (2005). 'Feminist legal theory'. *Journal of Gender, Social Policy and the Law* 13: 13.

Fredman, S (2001). Equality: A New Generation? *Industrial Law Journal* 30: 145–68.

Fredman, S. (2002). *Discrimination Law*. Oxford: Clarendon Press.

Fredman, S. (2011a). *Discrimination Law*, 2nd edn. Oxford: Oxford University Press.

Fredman, S. (2011b). 'The public sector equality duty'. *Industrial Law Journal* 40: 405–27.

Freedland, M. R. and Kountouris, N. (2008). 'Towards a comparative theory of the contractual construction of personal work relations in Europe'. *Industrial Law Journal* 37: 49–74.

Government Equalities Office (GEO) (2011). 'Addressing Gender Equality: "Think, Act, Report"'. GEO, London, September.

Green, T. K. (2003). 'Discrimination in workplace dynamics: Toward a structural account of disparate treatment theory'. *Harvard Civil Rights: Civil Liberties Law Review* 38: 91–157.

Greenwald, A. G. and Krieger, L. H. (2006). 'Implicit bias: Scientific foundations'. *California Law Review* 94: 945–67.

Hardin, C. D. and Banaji, M. R. (2013). 'The nature of implicit prejudice: Implications for personal and public policy'. In Shafir, E. (ed.) *The Behavioral Foundations of Public Policy.* Princeton, NJ: Princeton University Press: 13–31.

Hepple, B. (2008). 'The aims of equality law'. *Current Legal Problems* 61: 1–22.

Hepple, B. (2011). 'Enforcing equality law: Two steps forward and two steps backwards for reflexive regulation'. *Industrial Law Journal* 40: 315–35.

Kang, J. (2005). 'Trojan horses of race'. *Harvard Law Review* 118: 1489–593.

Kotz, D. M. (2009). 'The financial and economic crisis of 2008: A systemic crisis of neoliberal capitalism'. *Review of Radical Political Economics* 41: 305–17.

Lacey, N. (1987). 'Legislation against sex discrimination: Questions from a feminist perspective'. *Journal of Law and Society* 14: 411–21.

Lawrence, C. R. (1987). 'The id, the ego, and equal protection: Reckoning with unconscious racism'. *Stanford Law Review* 39: 317.

MacPherson, W. (1999). 'The Stephen Lawrence inquiry: Report of an inquiry by Sir William MacPherson of Cluny'. The Stationery Office, London, February.

McCrudden, C. (2007). 'Equality legislation and reflexive regulation: A response to the discrimination law review's consultative paper'. *Industrial Law Journal* 36: 255–66.

Ministry of Justice (2012). 'Employment tribunals and EAT statistics, 2011–2012: 1 April 2011 to 31 March 2012'. London, September.

Moreau, S. (2010). 'What is discrimination?' *Philosophy and Public Affairs* 38: 143–79.

Rönnmar, M. (2006). 'The managerial prerogative and the employee's obligation to work: Comparative perspectives on functional flexibility'. *Industrial Law Journal* 35: 56–74.

Shin, P. S. (2010). 'Liability for unconscious discrimination? A thought experiment in the theory of employment discrimination law'. *Hastings Law Journal* 62: 67–102.

Squires, J. (2005). 'Is mainstreaming transformative? Theorizing mainstreaming in the context of diversity and deliberation'. *Social Politics* 12: 366–88.

Steele, I. (2010). 'Sex discrimination and the material factor defence under the Equal Pay Act 1970 and the Equality Act 2010'. *Industrial Law Journal* 39: 264–74.

Thaler, R. H. and Sunstein, C. R. (2009). *Nudge: Improving Decisions about Health, Wealth and Happiness*. London: Penguin.

Villiers, C. (2010). 'Achieving gender balance in the boardroom: Is it time for legislative action in the UK?' *Legal Studies* 30: 533–57.

Yeung, K. (2012). 'Nudge as fudge'. *Modern Law Review* 75: 122–48.

Index of Names

Index of Subjects