Oxford Handbooks Online

Theory of Mind

Alvin I. Goldman

The Oxford Handbook of Philosophy of Cognitive Science
Edited by Eric Margolis, Richard Samuels, and Stephen P. Stich

Print Publication Date: Jan 2012 Subject: Philosophy, Philosophy of Mind
Online Publication Date: May 2012 DOI: 10.1093/oxfordhb/9780195309799.013.0017

Abstract and Keywords

The article provides an overview of 'Theory of Mind' (ToM) research, guided by two classifications. The first covers four competing approaches to mentalizing such as the theory-theory, modularity theory, rationality theory, and simulation theory. The second classification is the first-person/third-person contrast. Jerry Fodor claimed that commonsense psychology is so good at helping predict behavior that it is practically invisible. It works well because the intentional states it posits genuinely exist and possess the properties generally associated with them. The modularity model has two principal components. First, whereas the child-scientist approach claims that mentalizing utilizes domain-general cognitive equipment, the modularity approach posits one or more domain-specific modules, which use proprietary representations and computations for the mental domain. Second, the modularity approach holds that these modules are innate cognitive structures, which mature or come on line at preprogrammed stages and are not acquired through learning. The investigators concluded that autism impairs a domain-specific capacity dedicated to mentalizing. Gordon, Jane Heal, and Alvin Goldman explained simulation theory in such a way that mind readers simulate a target by trying to create similar mental states of their own as proxies or surrogates of those of the target. These initial pretend states are fed into the mind reader's own cognitive mechanisms to generate additional states, some of which are then imputed to the target.

Keywords: theory of mind, mentalizing, modularity theory, rationality theory, simulation theory

"THEORY of Mind" (ToM) refers to the cognitive capacity to attribute mental states to self and others. Other names for the same capacity include "commonsense psychology," "naŧve psychology," "folk psychology," "mindreading" and "mentalizing." Mental attributions are commonly made in both verbal and nonverbal forms. Virtually all language communities, it seems, have words or phrases to describe mental states, including perceptions, bodily feelings, emotional states, and propositional attitudes

(beliefs, desires, hopes, and intentions). People engaged in social life have many thoughts and beliefs about others' (and their own) mental states, even when they do not verbalize them.

In cognitive science the core question in this terrain is: How do people execute this cognitive capacity? How do they, or their cognitive systems, go about the task of forming beliefs or judgments about others' mental states, states that are not directly observable? Less frequently discussed in psychology is the question of how people self-ascribe mental states. Is the same method used for both first-person and third-person ascription, or entirely different methods? Other questions in the terrain include: How is the capacity for ToM acquired? What is the evolutionary story behind this capacity? What cognitive or neurocognitive architecture underpins ToM? Does it rely on the same mechanisms for thinking about objects in general, or does it employ dedicated, domain-specific mechanisms? How does it relate to other processes of social cognition, such as imitation or empathy?

This chapter provides an overview of ToM research, guided by two classifications. The first articulates four competing approaches to (third-person) mentalizing, viz., the theory-theory, modularity theory, rationality theory, and simulation theory. The second classification is the first-person/third-person contrast. The bulk of the discussion is directed at third-person mindreading, but the final section addresses self-attribution. Finally, our discussion provides representative coverage of the principal fields of investigators of ToM: philosophy of mind, developmental (p. 403) psychology, and cognitive neuroscience. Each of these fields has its distinctive research style, central preoccupations, and striking discoveries or insights.

1. The Theory-Theory

Philosophers began work on ToM, or folk psychology, well before empirical researchers were seriously involved, and their ideas influenced empirical research. In hindsight one might say that the philosopher Wilfrid Sellars (1956) jump-started the field with his seminal essay, "Empiricism and the Philosophy of Mind." He speculated that the commonsense concepts and language of mental states, especially the propositional attitudes, are products of a proto-scientific *theory* invented by one of our fictional ancestors. This was the forerunner of what was later called the "theory-theory." This idea has been warmly embraced by many developmental psychologists. However, not everyone agrees with theory-theory as an account of commonsense psychology, so it is preferable to avoid the biased label "theory of mind." In much of my discussion, therefore, I opt for more neutral phraseology, "mindreading" or "mentalizing," to refer to the activity or trait in question.

The popularity of the theory-theory in philosophy of mind is reflected in the diversity of philosophers who advocate it. Jerry Fodor (1987) claims that commonsense psychology is so good at helping us predict behavior that it is practically invisible. It works well because the intentional states it posits genuinely exist and possess the properties generally associated with them. In contrast to Fodor's intentional realism, Paul Churchland (1981) holds that commonsense psychology is a radically false theory, one that ultimately should be eliminated. Despite their sharp differences, these philosophers share the assumption that naŧve psychology, at bottom, is driven by a science-like theory, where a theory is understood as a set of law-like generalizations. Naŧve psychology would include generalizations that link (1) observable inputs to certain mental states, (2) certain mental states to other mental states, and (3) mental states to observable outputs (behavior). The first type of law might be illustrated by “Persons who have been physically active without drinking fluids tend to feel thirst.” An example of the second might be “Persons in pain tend to want to relieve that pain.” An example of the third might be “People who are angry tend to frown.” The business of attributing mental states to others consists of drawing law-guided inferences from their observed behavior, stimulus conditions, and previously determined antecedent mental states. For example, if one knows that Melissa has been engaged in vigorous exercise without drinking, one may infer that she is thirsty.

Among the developmental psychologists who have championed the theory-theory are Josef Perner, Alison Gopnik, Henry Wellman, and Andrew Meltzoff. They seek to apply it to young children, who are viewed as little scientists who form and revise their thinking about various domains in the same way scientists do (Gopnik and Wellman 1992; Gopnik and Meltzoff 1997). They collect evidence, make observations, and change their theories in a highly science-like fashion. They generate theories not only about physical phenomena but about unobservable mental states such as belief and desire. As in formal science, children make transitions from simple theories of the phenomena to more complex ones.

The most famous empirical discovery in the developmental branch of theory of mind is the discovery by Wimmer and Perner (1983) of a striking cognitive change in children between roughly three and four years of age. This empirical discovery is that three-year-olds tend to fail a certain *false-belief task* whereas four-year-olds tend to succeed on the task. Children watch a scenario featuring puppets or dolls in which the protagonist, Sally, leaves a chocolate on the counter and then departs the scene. In her absence Anne is seen to move the object from the counter to a box. The children are asked to predict where Sally will look for the chocolate when she returns to the room, or alternatively where Sally “thinks” the chocolate is. Prior to age four children typically answer incorrectly, that is, that Sally thinks it is in the box (where the chocolate really is). Around age four, however, normal children answer as an adult would, by specifying the place where Sally left the chocolate, thereby ascribing to Sally (what they recognize to be) a

false belief. What happens between three and four that accounts for this striking difference?

Theory-theorists answer by positing a change of theory in the minds of the children. At age three they typically have conceptions of desire and belief that depict these states as simple relations between the cognizer and the external world, relations that do not admit the possibility of error. This simple theory gradually gives way to a more sophisticated one in which beliefs are related to propositional representations that can be true or false of the world. At age three the child does not yet grasp the idea that a belief can be false. In lacking a representational theory of belief, the child has—as compared with adults—a "conceptual deficit" (Perner 1991). This deficit is what makes the three-year-old child incapable of passing the false-belief test. Once the child attains a representational theory of belief, roughly at age four, she passes the location-change false-belief test.

A similar discrepancy between three- and four-year olds was found in a second type of false-belief task, the deceptive container task. A child is shown a familiar container that usually holds candy and is asked, "What's in here?" He replies, "candy." The container is then opened, revealing only a pencil. Shortly thereafter the child is asked what he thought was in the container when he was first asked. Three-year-olds incorrectly answer "a pencil," whereas four-year-olds correctly answer "candy." Why the difference between the two age groups, despite the fact that memory tests indicate that three-year-olds have no trouble recalling their own psychological states? Theory-theorists again offered the same conceptual-deficit explanation. Since the three-year-olds' theory does not leave room for the possibility of false belief, they cannot ascribe to themselves their original (false) belief that the container held candy, so they respond with their current belief, namely, that it held a pencil.

(p. 405) This explanation was extremely popular circa 1990. But several subsequent findings seriously challenge the conceptual-deficit approach. The early challenges were demonstrations that various experimental manipulations enable three-year-olds to pass the tests. When given a memory aid, for example, they can recall and report their original false prediction (Mitchell and Lacohee 1991). They can also give the correct false-belief answer when the reality is made less salient, for instance, if they are told where the chocolate is but do not see it for themselves (Zaitchik 1991). Additional evidence suggests that the problem with three-year-olds lies in the area of inhibitory control (Carlson and Moses 2001). Inhibitory control is an executive ability that enables someone to override "prepotent" tendencies (i.e., dominant or habitual tendencies, such as the tendency to reference reality as one knows it to be). A false-belief task requires an attributor to override this natural tendency, which may be hard for three-year-olds. An extra year during which the executive powers mature may be the crucial difference for four-year-olds, not a change in their belief concept. A meta-analysis of false-belief task findings encourages Wellman, Cross, and Watson (2001) to retain the conceptual-deficit story, but this is strongly disputed by Scholl and Leslie (2001).

Even stronger evidence against the traditional theory-theory time line was uncovered in 2005, in a study of fifteen-month-old children using a nonverbal false-belief task. Onishi and Baillargeon (2005) employed a new paradigm with reduced task demands to probe the possible appreciation of false belief in fifteen-month-old children, and found signs of exactly such understanding. This supports a *much* earlier picture of belief understanding than the child-scientist form of theory-theory ever contemplated.

A final worry about this approach can now be added. A notable feature of professional science is the diversity of theories that are endorsed by different practitioners. Cutting-edge science is rife with disputes over which theory to accept—disputes that often persist for decades. This pattern of controversy contrasts sharply with what is ascribed to young children in the mentalizing domain. They are said to converge on one and the same theory, all within the same narrow time-course. This bears little resemblance to professional science.

Gopnik takes a somewhat different tack in recent research. She puts more flesh on the general approach by embedding it in the Bayes-net formalism. Bayes nets are directed-graph formalisms designed to depict probabilistic causal relationships between variables. Given certain assumptions (the causal Markov and faithfulness assumptions), a system can construct algorithms to arrive at a correct Bayes net causal structure if it is given enough information about the contingencies or correlations among the target events. Thus, these systems can learn about causal structure from observations and behavioral interventions. Gopnik and colleagues (Gopnik et al. 2004; Schulz and Gopnik 2004) report experimental results suggesting that two- to four-year-old children engage in causal learning in a manner consistent with the Bayes net formalism. They propose that this is the method used to learn causal relationships between mental variables, including relationships relevant to false-belief tasks.

(p. 406) Here are several worries about this approach. Can the Bayes net formalism achieve these results without special tweaking by the theorist, and if not, can other formalisms match these results without similar “special handling”? Second, if the Bayes-net formalism predicts that normal children make all the same types of causal inferences, does this fit the scientific inference paradigm? We again encounter the problem that scientific inference is characterized by substantial diversity across the community of inquirers, whereas the opposite is found in the acquisition of mentalizing skills.

2. The Modularity-Nativist Approach to Theory of Mind

In the mid-1980s, other investigators found evidence supporting a very different model of ToM acquisition. This is the *modularity* model, which has two principal components. First, whereas the child-scientist approach claims that mentalizing utilizes domain-general cognitive equipment, the modularity approach posits one or more domain-specific modules, which use proprietary representations and computations for the mental domain. Second, the modularity approach holds that these modules are innate cognitive structures, which mature or come on line at preprogrammed stages and are not acquired through learning (Leslie 1994; Scholl and Leslie 1999). This approach comports with nativism for other domains of knowledge, such as those subsumed under Spelke's (1994) idea of “core knowledge.” The core-knowledge proposal holds that infants only a few months old have a substantial amount of “initial” knowledge in domains such as physics and arithmetic: knowledge that objects must trace spatiotemporally continuous paths through space, or that one plus one yields two. Innate principles are at work that are largely independent of and encapsulated from one another. Modularists about mentalizing endorse the same idea. Mentalizing is part of our genetic endowment that is triggered by appropriate environmental factors, just as puberty is triggered rather than learned (Scholl and Leslie 2001).

Early evidence in support of a psychology module was reported by Simon Baron-Cohen, Alan Leslie, and Uta Frith in two studies, both concerning autism. The first (Baron-Cohen et al. 1985) compared the performance of normal preschool children, Down syndrome children, and autistic children on a false-belief task. All children had a mental age of above four years, although the chronological age of the second two groups was higher. Eighty-five percent of the normal children, 86 percent of the Down syndrome children, but only 20 percent of the autistic children passed the test. In the second study (1986) subjects were given scrambled pictures from comic strips with the first picture already in place. They were supposed to put the strips in order to make a coherent story, and were also supposed to tell the story in their own words. The stories were of three types: mechanical, behavioral, and (p. 407) mentalistic. The autistic children all ordered the mechanical strips correctly and dealt adequately with the behavioral script. But the vast majority of autistic children could not understand the mentalistic stories. They put the pictures in jumbled order and told stories without attribution of mental states.

The investigators concluded that autism impairs a domain-specific capacity dedicated to mentalizing. Notice that the autistic children in the 1986 study were not deficient on either the mechanical or the behavioral script, only on the mentalistic one. Conversely, the Down syndrome children, despite their general retardation, were not deficient on the false-belief task. Thus autism seems to involve an impairment specific to mentalizing,

whereas mentalizing need not be impaired by general retardation as long as the ToM-dedicated module remains intact.

These conclusions, however, are not entirely secure. Some children with autism pass ToM tasks, including false-belief tests. The number who pass varies from one study to the next, but even a small percentage calls for explanation. If autism involves a failure to develop a ToM, how could these participants with autism pass the tests? Others therefore argue that failure on tasks that tap mentalizing abilities may be more directly interpreted in terms of domain-general deficits in either executive functions or language (Tager-Flusberg 2000).

Nativist modularists adduce additional evidence, however, in support of their view, especially evidence for an appreciation of intentional agency in preverbal infants. A variety of cues are cited as evidence for the attribution of intentionality, or goal-directedness, in infancy, including joint attention behaviors (gaze-following, pointing, and other communicative gestures), imitation, language and emotional referencing, and looking-time studies.

In one study of gaze-following, Johnson, Slaughter, and Carey (1998) tested twelve-month-old infants on a novel object, a small, beach ball-sized object with natural-looking fuzzy brown fur. It was possible to control the object's behavior from a hidden vantage point so that when the baby babbled, the object babbled back. After a period of familiarization, an infant either experienced the object reacting contingently to the infant's own behavior or merely random beeping or flashing. Infants followed the “gaze” of the object by shifting their own attention in the same direction under three conditions: if the object had a face, or if the object beeped and flashed contingent on the infant's own behavior, or both. These results were interpreted as showing that infants use specific information to decide when an object does or does not have the ability to perceive or attend to its surroundings, which seems to support the operation of a dedicated input system (Johnson 2005). Woodward (1998) used a looking-time measure to show that even five-month-olds appear to interpret human hands as goal-directed relative to comparable inanimate objects. They looked longer if the goal-object of the hand changed, but not if the hand's approach path to the goal-object changed. This evidence also suggests an early, dedicated system to the detection of goal-oriented entities.

All of the above findings post-date Alan Leslie's (1994) postulation of a later-maturing cognitive module: the “theory-of-mind mechanism” (ToMM). Leslie highlighted **(p. 408)** four features of ToMM: (1) it is domain specific, (2) it employs a proprietary representational system that describes propositional attitudes, (3) it forms the innate basis for our capacity to acquire ToM, and (4) it is damaged in autism. ToMM uses specialized representations and computations; it is fast, mandatory, domain specific, and informationally encapsulated, thereby satisfying the principal characteristics of modularity as described by Fodor (1983).

An initial problem with the modularity theory is that ToMM, the most widely discussed module postulated by the theory, does not satisfy the principal criteria of modularity associated with Fodorian modularity. Consider domain specificity. Fodor says that a cognitive system is domain specific just in the case when “only a restricted class of stimulations can throw the switch that turns [the system] on” (1983, 49). It is doubtful that any suitable class of stimulations would satisfy this condition for ToMM (Goldman 2006, 102–4). A fundamental obstacle facing this proposal, moreover, is that Fodor's approach to modularity assumes that modules are either input systems or output systems, whereas mindreading has to be a central system. Next consider informational encapsulation, considered the heart of modularity. A system is informationally encapsulated if it has only limited access to information contained in other mental systems. But when Leslie gets around to illustrate the workings of ToMM, it turns out that information from other central systems is readily accessible to ToMM (Nichols and Stich, 2003, 117–21). Leslie and German (1995) discuss an example of ascribing a pretend state to another person, and clearly indicate that a system ascribing such a pretense uses real-world knowledge, for example, whether a cup containing water would disgorge its contents if it were upturned. This knowledge would have to be obtained from (another) central system. Perhaps such problems can be averted if a non-Fodorian conception of modularity is invoked, as proposed by Carruthers (2006). But the tenability of the proposed alternative conception is open to debate.

3. The Rationality-Teleology Theory

A somewhat different approach to folk psychology has been championed by another group of philosophers, chief among them Daniel Dennett (1987). Their leading idea is that one mind reads a target by “rationalizing” her, that is, by assigning to her a set of propositional attitudes that make her emerge—as far as possible—as a rational agent and thinker. Dennett writes:

> [I]t is the myth of our rational agenthood that structures and organizes our attributions of belief and desire to others and that regulates our own deliberations and investigations…. Folk psychology, then, is *idealized* in that it produces its predictions and explanations by calculating in a normative system; it predicts what we will believe, desire, and do, by determining what we ought to believe, desire, and do. (1987, 52)

(p. 409) Dennett contends that commonsense psychology is the product of a special stance we take when trying to predict others’ behavior: the *intentional stance.* To adopt the intentional stance is to make the default assumption that the agent whose behavior is to be predicted is rational, that her desires and beliefs, for example, are ones she rationally ought to have given her environment and her other beliefs or desires.

Dennett does not support his intentional stance theory with empirical findings; he proceeds largely by thought experiment. So let us use the same procedure in evaluating his theory. One widely endorsed normative principle of reasoning is to believe whatever follows logically from other things we believe. But attributors surely do not predict their targets' belief states in accordance with such a strong principle; they do not impute "deductive closure" to them. They allow for the possibility that people forget or ignore many of their prior beliefs and fail to draw all of the logical consequences that might be warranted (Stich 1981). What about a normative rule of inconsistency avoidance? Do attributors assume that their targets conform to this requirement of rationality? That too seems unlikely. If an author modestly thinks that he must have made some error in his book packed with factual claims, he is caught in an inconsistency (this is the so-called "paradox of the preface"). But would attributors not be willing to ascribe belief in all these propositions to this author?

These are examples of implausible consequences of the rationality theory. A different problem is the theory's incompleteness: it covers only the mindreading of propositional attitudes. What about other types of mental states, such as sensations like thirst or pain and emotions like anger or happiness? It is dubious that rationality considerations bear on these kinds of states, yet they are surely among the states that attributers ascribe to others. There must be more to mindreading than imputed rationality.

Although first inspired by armchair reflection, rationality theory has also inspired some experimental work that—at least at first blush—seems to be supportive. Gergely et al. (1995) performed an intriguing experiment they interpreted as showing that toddlers take the intentional stance at twelve months of age. They habituated one-year-old infants to an event in which a small circle approaches a large circle by jumping over an obstacle. When the obstacle is later removed, the infants show longer looking-times when they see the circle take the familiar jumping path as compared with a straight path toward the target. Apparently, infants expect an agent to take the most rational or efficient means to its goal, so they are surprised when it takes the jumping path, although that is what they have seen it do in the past.

The title of Gergely et al.'s (1995) paper, "Taking the Intentional Stance at 12 Months of Age," conveyed the influence of Dennett's rationality theory. The authors' first interpretation of the results articulated this theme, viz., that infants attribute a causal intention to the agent that accords with a rationality principle. Toward the end of their paper, however, they concede that an infant can represent the agent's action as intentional without attributing a mental representation of the future goal state to the agent's mind. Thus, the findings might simply indicate that (p. 410) the infant represents actions by relating relevant aspects of reality (action, goal-state, and situational constraints) through a principle of efficient action, which assumes that actions function to realize goal-states by the most efficient means available. Indeed, in subsequent writings the authors switch their description of infants from the "intentional" stance to the "teleological" stance, an interpretational system for actions in terms of means-ends efficiency (Gergely and Csibra 2003). The teleological stance is a qualitatively different

but developmentally related interpretational system that is supposed to be the precursor of the young child's intentional stance. The two stances differ in that teleological interpretation is nonmentalistic—it makes reference only to actual and future states of reality. Developmentally, however, teleological interpretation is transformed into causal mentalistic interpretation by "mentalizing" the explanatory constructs of the teleological stance (232).

This approach raises three problems. First, can the teleological stance really be transformed into the full range of mentalistic interpretation in terms of rationality principles? One species of mindreading involves imputing beliefs to a target based on inferential relations to prior belief states. How could this interpretational system be a transformation of an efficiency principle? Inference involves no action or causal efficiency. Second, the teleological stance might equally be explained by a rival approach to mentalizing, namely, the simulation theory. The simulation theory might say that young children project themselves into the shoes of the acting object (even a circle) and consider the most efficient means to its goal. They then expect the object to adopt this means. Third, as already noted above, there are kinds of mental states and mindreading contexts that have nothing to do with rationality or efficiency. People ascribe emotional states to others (fear or delight, disgust or anger) based on facial expressions. How could these ascriptions be driven by a principle of efficiency? We do not have the makings here of a general account of mindreading, but rather at most, a narrow segment of it. And even this narrow segment might be handled just as well by a rival theory (viz., the simulation theory).

4. The Simulation Theory

A fourth approach to commonsense psychology is the *simulation theory*, sometimes called the "empathy theory." Robert Gordon (1986) was the first to develop this theory in the present era, suggesting that we can predict others' behavior by answering the question, "What would *I* do in *that* person's situation?" Chess players playing against a human opponent report that they visualize the board from the other side, taking the opposing pieces for their own and vice versa. They pretend that their reasons for action have shifted accordingly. Thus transported in imagination, they make up their mind what to do and project this decision onto the opponent.

The basic idea of the simulation theory resurrects ideas from a number of earlier European writers, especially in the hermeneutic tradition. Dilthey wrote of (p. 411) understanding others through a process of "feeling with" others (*mitfuehlen*), "reexperiencing" (*nacherleben)* their mental states, or "putting oneself into" (*hineinversetzen)* their shoes. Similarly, Schleiermacher linked our ability to understand other minds with our capacity to imaginatively occupy another person's point of view. In the philosophy of history, the English philosopher R. G. Collingwood (1946)

suggested that the inner imitation of thoughts, or what he calls the reenactment of thoughts, is a central epistemic tool for understanding other agents. (For an overview of this tradition, see Stueber 2006.)

In addition to Gordon, Jane Heal (1986) and Alvin Goldman (1989) in the 1980s endorsed the simulation idea. Their core idea is that mind readers simulate a target by trying to create similar mental states of their own as proxies or surrogates of those of the target. These initial pretend states are fed into the mind reader's own cognitive mechanisms to generate additional states, some of which are then imputed to the target. In other words, attributors use their own mind to mimic or “model” the target's mind and thereby determine what has or will transpire in the target.

An initial worry about the simulation idea is that it might “collapse” into theory theory. As Dennett put the problem:

> How can [the idea] work without being a kind of theorizing in the end? For the state I put myself in is not belief but make-believe belief. If I make believe I am a suspension bridge and wonder what I will do when the wind blows, what “comes to me” in my make-believe state depends on how sophisticated my knowledge is of the physics and engineering of suspension bridges. Why should my making believe I have your beliefs be any different? In both cases, knowledge of the imitated object is needed to drive the make-believe “simulation,” and the knowledge must be organized into something rather like a theory. (1987, 100–101)

Goldman (1989) responded that there is a difference between *theory-driven* simulation, which must be used for systems different than oneself, and *process-driven* simulation, which can be applied to systems resembling oneself. If the process or mechanism driving the simulation is similar enough to the process or mechanism driving the target, and if the initial states are also sufficiently similar, the simulation might produce an isomorphic final state to that of the target without the help of theorizing.

5. Mirroring and Simulational Mindreading

The original form of simulation theory (ST) primarily addressed the attribution of propositional attitudes. In recent years, however, ST has focused heavily on simpler mental states, and on processes of attribution rarely dealt with in the early ToM literature. We include here the mindreading of motor plans, sensations, and emotions. This turn in ST dates to a paper by Vittorio Gallese and Alvin Goldman (1998), **(p. 412)** which posited a link between simulation-style mindreading and activity of mirror neurons (or mirror systems). Investigators in Parma, Italy, led by Giacomo Rizzolatti, first discovered mirror neurons in macaque monkeys by using single cell recordings (Rizzolatti et al. 1996; Gallese et al. 1996). Neurons in the macaque premotor cortex often code for a particular type of goal-oriented action, for example, grasping, tearing, or manipulating an

object. A subclass of premotor neurons were found to fire both when the animal plans to perform an instance of its distinctive type of action and when it observes another animal (or human) perform the same action. These neurons were dubbed “mirror neurons,” because an action plan in the actor's brain is mirrored by a similar action plan in the observer's brain. Evidence for a mirror system in humans was established around the same time (Fadiga et al. 1995). Since the mirror system of an observer tracks the mental state (or brain state) of an agent, the observer executes a mental simulation of the latter. If this simulation also generates a mental-state attribution, this would qualify as simulation-based mindreading. It would be a case in which an attributor uses his own mind to “model” that of the target. Gallese and Goldman speculated that the mirror system might be part of, or a precursor to, a general mindreading system that works on simulationist principles.

Since the mid-1990s the new discoveries of mirror processes and mirror systems have expanded remarkably. Motor mirroring has been established via sound as well as vision (Kohler et al. 2002), and for effectors other the hand, specifically, the foot and the mouth (Buccino et al. 2001). Meanwhile, mirroring has been discovered for sensations and emotions. Under the category of sensations, there is mirroring for touch and mirroring for pain. Touching a subject's legs activates primary and secondary somatosensory cortex. Keysers et al. (2004) showed subjects movies of other subjects being touched on their legs. Large extents of the observer's somatosensory cortex also responded to the sight of the targets’ legs being touched. Several studies established mirroring for pain in the same year (Singer et al. 2004, Jackson et al. 2004, and Morrison et al. 2004). In the category of emotions, the clearest case is mirroring for disgust. The anterior insula is well-known as the primary brain region associated with disgust. Wicker et al. (2003) undertook an fMRI experiment in which normal subjects were scanned while inhaling odorants through a mask—either foul, pleasant, or neutral—and also while observing video clips of other people's facial expressions while inhaling such odorants. Voxels in the anterior insula that were significantly activated when a person inhaled a foul odorant were also significantly activated when seeing others make facial expressions arising from a foul odorant. Thus, there was mirroring of disgust.

The critical question for ToM, however, is whether mindreading (i.e., mental attribution) occurs as an upshot of mirroring. In 2005 two similar experiments in the domain of motor intention were performed by members of the Parma group, and are claimed to provide evidence for mirror-based—hence, simulation-based—prediction of motor intentions. One experiment was done with monkeys (Fogassi et al. 2005) and the other with humans (Iacoboni et al. 2005). This section shall sketch the latter study only.

(p. 413) Iacoboni et al.'s study was an fMRI study in which subjects observed video clips presenting three kinds of stimulus conditions: (1) grasping hand actions without any context (“Action” condition), (2) scenes specifying a context without actions (i.e., a table set for drinking tea or ready to be cleaned up after tea) (“Context” condition), and (3) grasping hand actions performed in either the before-tea or the after-tea context (“Intention” condition). The Intention condition yielded a significant signal increase in

premotor mirroring areas where hand actions are represented. The investigators interpreted this as evidence that premotor mirror areas are involved in understanding the intentions of others, in particular, intentions to perform subsequent actions (e.g., drinking tea or cleaning up).

This mindreading conclusion, however, is somewhat problematic, because there are alternative "deflationary" interpretations of the findings (Goldman 2008). One deflationary interpretation would say that the enhanced activity in mirror neuron areas during observation of the Intention condition involved only predictions of *actions*, not attributions of *intentions.* Since actions are not mental states, predicting actions does not qualify as mindreading. The second deflationary interpretation is that the activity in the observer's relevant mirror area is a mimicking of the agent's intention, not an intention *attribution* (belief). Reexperiencing an intention should not be confused with attributing an intention. Only the attribution of an intention would constitute a belief or judgment about an intention. Thus, the imaging data do not conclusively show that mindreading took place in the identified premotor area.

However, the Iacoboni et al. study presented evidence of intention attribution above and beyond the fMRI evidence. After being scanned, subjects were debriefed about the grasping actions they had witnessed. They all reported representing the intention of drinking when seeing the grasping action in the during-tea condition and representing the intention of cleaning up when seeing the grasping action in the after-tea condition. Their verbal reports were independent of the instructions the subjects had been given at the outset. Thus, it is quite plausible that their reported intention attributions were caused by activity in the mirror area. So the Iacoboni et al. study does provide positive evidence for its stated conclusion, even if the evidence is not quite as probative as its researchers contend.

Where else might we look for evidence of mirroring-based mindreading? Better specimens of evidence are found in the emotion and sensation domains. For reasons of space, attention is restricted here to emotion. Although Wicker et al. (2003) established a mirror process for disgust, they did not test for disgust attribution. However, by combining their fMRI study of normal subjects with neuropsychological studies of brain-damaged patients, a persuasive case can be made for mirror-caused disgust attribution (in normals). Calder et al. (2000) studied patient NK, who suffered insula and basal ganglia damage. In questionnaire responses NK showed himself to be selectively impaired in experiencing disgust, as contrasted with fear or anger. NK also showed significant and selective impairment in disgust recognition (attribution), in both visual and auditory modalities. Similarly, Adolphs et al. (2003) had a patient B who suffered extensive damage to the anterior **(p. 414)** insula but was able to recognize the six basic emotions *except disgust* when observing dynamic displays of facial expressions. The inability of these two patients to undergo a normal disgust response in their anterior insula apparently prevented them from mindreading disgust in others, although their attribution of other basic emotions was preserved. It is reasonable to conclude that when

normal individuals recognize disgust through facial expressions of a target, this is causally mediated by a mirrored experience of disgust (Goldman and Sripada 2005; Goldman 2006).

Low-level mindreading, then, can be viewed as an elaboration of a primitive tendency to engage in automatic mental mimicry. Both behavioral and mental mimicry are fundamental dimensions of social cognition. Meltzoff and Moore (1983) found facial mimicry in neonates less than an hour old. Among adults, unconscious mimicry in social situations occurs for facial expressions, hand gestures, body postures, speech patterns, and breathing patterns (Hatfield, Cacioppo, and Rapson 1994; Bavelas et al. 1986; Dimberg, Thunberg, and Elmehed 2000; Paccalin and Jeannerod 2000). Chartrand and Bargh (1999) found that automatic mimicry occurs even between strangers, and that it leads to higher liking and rapport between interacting partners. Mirroring, of course, is mental mimicry usually unaccompanied by behavioral mimicry. The sparseness of behavioral imitation (relative to the amount of mental mimicry) seems to be the product of inhibition. Compulsive behavioral imitation has been found among patients with frontal lesions, who apparently suffer from an impairment of inhibitory control (Lhermitte et al. 1986; de Renzi et al. 1996). Without the usual inhibitory control, mental mimicry would produce an even larger amount of behavioral mimicry. Thus, mental mimicry is a deep-seated property of the social brain, and low-level mindreading builds on its foundation.

6. Simulation and High-Level Mindreading

The great bulk of mindreading, however, cannot be explained by mirroring. Can it be explained (in whole or part) by another form of simulation? The general idea of mental simulation is the reexperiencing or reenactment of a mental event or process, or an *attempt* to reexperience or reenact a mental event (Goldman 2006, ch. 2). Where does the traditional version of simulation theory fit into the picture? It mainly fits into the second category, that is, *attempted* interpersonal reenactment. This captures the idea of mental pretense, or what we call "enactment imagination" (E-imagination), which consists of trying to construct in oneself a mental state that is not generated by the usual means (Goldman 2006; Currie and Ravenscroft 2002). *Simulating Minds* argues that E-imagination is an intensively used cognitive operation, one commonly used in reading others' minds.

Let us first illustrate E-imagination with intrapersonal applications, for example, imagining seeing something or launching a bodily action. The products of such (p. 415) applications constitute, respectively, visual and motor imagery. To visualize something is to (try to) construct a visual image that resembles the visual experience we would undergo if we were actually seeing what is visualized. To visualize the *Mona Lisa* is to (try to) produce a state that resembles a seeing of the *Mona Lisa*. Can visualizing really resemble vision? Cognitive science and neuroscience suggest an affirmative answer. Kosslyn (1994) and others have shown how the processes and products of visual perception and visual imagery have substantial overlap. An imagined object "overflows" the visual field of imagination at about the same imagined distance from the object as it overflows the real visual field. This was shown in experiments where subjects actually walked toward rectangles mounted on a wall and when they merely visualized the rectangles while imagining a similar walk (Kosslyn 1978). Neuroimaging reveals a notable overlap between parts of the brain active during vision and during imagery. A region of the occipitotemporal cortex known as the fusiform gyrus is activated both when we see faces and when we imagine them (Kanwisher et al. 1997). Lesions of the fusiform face area impair both face recognition and the ability to imagine faces (Damasio et al. 1990).

An equally (if not more) impressive story can be told for motor imagery. Motor imagery occurs when you are asked to imagine (from a motoric perspective) moving your effectors in a specified way, for example, playing a piano chord with your left hand or kicking a soccer ball. It has been shown convincingly that motor imagery corresponds closely, in neurological terms, to what transpires when you actually execute the relevant movements (Jeannerod 2001).

At least in some modalities, then, E-imagination produces strikingly similar experiences to ones that are usually produced otherwise. Does the same hold for mental events such as forming a belief or making a decision? This has not been established, but it is entirely consistent with existing evidence. Moreover, a core brain network has recently been proposed that might underpin high-level simulational mindreading as a special case.

Buckner and Carroll (2007) propose a brain system that subserves at least three, and possibly four, forms of what they call “self-projection.” Self-projection is the projection of the current self into one's personal past or one's personal future, and also the projection of oneself into other people's minds or other places (as in navigation). What all these mental activities share is projection of the self into alternative situations, involving a perspective shift from the immediate environment to an imagined environment (the past, the future, other places, other minds). Buckner and Carroll refer to the mental construction of an imagined alternative perspective as a “simulation.”

So E-imaginative simulation might be used successfully for reading other minds. But what specific evidence suggests that we deploy E-imaginative simulation in trying to mindread others, much of the time or even most of the time? This is what simulation theory concerning high-level mindreading needs to establish. (This assumes that simulation theory no longer claims that each and every act of mindreading is executed by simulation. Rather, simulation theorists are prepared to accept a hybrid approach in which simulation plays a central but not exclusive role.)

(p. 416) Two lines of evidence will be presented here (for additional lines of argument, see Goldman 2006, ch. 7). An important feature of the imagination-based simulation story is that successful mindreading requires a carefully pruned set of pretend inputs in the simulational exercise. The exercise must not only *include* pretend or surrogate states that correspond to those of the target but also *exclude* the mindreader's own genuine states that do not correspond to ones of the target. This implies the possibility of two kinds of error or failure: failure to include states possessed by the target, and failure to exclude states lacked by the target. The second type of error will occur if a mindreader allows a genuine state of his own, which he “knows” that the target lacks, to creep into the simulation and contaminate it. This is called *quarantine failure*. There is strong evidence that quarantine failure is a serious problem for mental-state attributors. This supports ST because quarantine failure is a likely affliction if mindreading is executed by simulation, but should pose no comparable threat if mindreading is executed by theorizing.

Why is it a likely problem under the simulation story? If one tries to predict someone's decision via simulation, one sets oneself to make a decision (in pretend mode). In making this decision, one's own relevant desires and beliefs try to enter the field to “throw their weight around” because this is their normal job. It is difficult to monitor the states that do not belong there, however, and enforce their departure. Enforcement requires suppression or inhibition, which takes vigilance and effort. No analogous problem rears its head under a theorizing scenario. If theorizing is used to predict a target's decision, an attributor engages in purely factual reasoning, not in mock decision making, so there is no reason why his genuine first-order desires or beliefs should intrude. What matters to the factual reasoning are the mindreader's beliefs *about* the target's desires and beliefs, and these second-order beliefs pose no comparable threat of intrusion.

Evidence shows that quarantine failure is in fact rampant, a phenomenon generally known as “egocentric bias.” Egocentric biases have been found for knowledge, valuation, and feeling. In the case of knowledge, egocentric bias has been labeled “the curse of knowledge,” and it has been found in both children (Birch and Bloom 2003) and adults (Camerer et al. 1989). To illustrate the bias for valuations, Van Boven, Dunning, and Loewenstein (2000) gave subjects Cornell coffee mugs and then asked them to indicate the lowest price they would sell their mugs for, while others who did not receive mugs were asked to indicate the highest price they would pay to purchase one. Because prices reflect valuations, the price estimates were, in effect, mental-state predictions. Both owners and sellers substantially underestimated the differences in valuations between themselves and their opposite numbers, apparently projecting their own valuations onto others. This gap proved very difficult to eliminate. To illustrate the case of feelings, Van Boven and Loewenstein (2003) asked subjects to predict the feelings of hikers lost in the woods with neither food nor water. What would bother them more, hunger or thirst? Predictions were elicited either before or after the subjects engaged in vigorous exercise, which would make one thirsty. Subjects who had just exercised were more likely to predict that (p. 417) the hikers would be more bothered by thirst than by hunger, apparently allowing their own thirst to contaminate their predictions.

Additional evidence that effective quarantine is crucial for successful third-person mindreading comes from neuropsychology. Samson et al. (2005) report the case of patient WBA, who suffered a lesion to the right inferior and middle frontal gyri. His brain lesion included a region previously identified as sustaining the ability to inhibit one's own perspective. Indeed, WBA had great difficulty precisely in inhibiting his own perspective (his own knowledge, desires, emotions, etc.). In nonverbal false-belief tests, WBA made errors in eleven out of twelve trials where he had to inhibit his own knowledge of reality. Similarly, when asked questions about other people's emotions and desires, which again required him to inhibit his own perspective, fifteen of twenty-seven responses involved egocentric errors. This again supports the simulationist approach to high-level mindreading. There is, of course, a great deal of other relevant evidence, which requires considerable interpretation and analysis. But ST seems to fare well in light of recent evidence (for contrary assessments, see Saxe 2005 and Carruthers 2006).

7. First-Person Mindreading

Our last topic is self-mentalization. Philosophers have long claimed that a special method —“introspection,” or “inner sense”—is available for detecting one's own mental states, although this traditional view is the object of skepticism and even scorn among many scientifically minded philosophers and cognitive scientists. Most theory theorists and rationality theorists would join these groups in rejecting so-called “privileged access” to one's own current mental states. Theory theorists would say that self-ascription, like

other-person ascription, proceeds by theoretical inference (Gopnik 1993). Dennett holds that the intentional stance is applied even to oneself. But these positions can be challenged with simple thought experiments, such as this one:

> I am now going to predict my bodily action during the next twenty seconds. It will include, first, curling my right index finger, then wrinkling my nose, and finally removing my glasses. There, those predictions are verified! I did all three things. You could not have duplicated these predictions (with respect to *my* actions). How did I manage it? Well, I let certain intentions form, and then I detected (i.e., introspected) those intentions. The predictions were based on the introspections. No other clues were available to me, in particular, no behavioral or environmental cues. The predictions must have been based, then, on a distinctive form of access I possess vis-à-vis my current states of mind, in this case, states that were primed to cause the actions. I seem to have similar access to my own itches and memories.

In an important modification of a well-known paper that challenged the existence or reliability of introspective access (Nisbett and Wilson 1977), the coauthor Wilson **(p. 418)** subsequently provides a good example and a theoretical correction to the earlier paper:

> The fact that people make errors about the causes of their own responses does not mean that their inner worlds are a black box. I can bring to mind a great deal of information that is inaccessible to anyone but me. Unless you can read my mind, there is no way you could know that a specific memory just came to mind, namely an incident in high school in which I dropped my bag lunch out a third-floor window, narrowly missing a gym teacher …. Isn't this a case of my having privileged, "introspective access to higher order cognitive processes"? (2002, 105)

Nonetheless, developmentalists have adduced evidence that putatively supports a symmetry or parallelism between self and other. They deny the existence of a special method, or form of access, available only to the first-person. Nichols and Stich (2003, 168–92) provide a comprehensive analysis of this literature, with the clear conclusion that the putative parallelism does not hold up, and fails precisely in ways that favor introspection or self-monitoring.

If there is such a special method, how exactly might it work? Nichols and Stich present their own model of self-monitoring. To have beliefs about one's own beliefs, they say, all that is required is that there be a monitoring mechanism that, when activated, takes the representation *p* in the Belief Box as input and produces the representation *I believe that p* as output. To produce representations of one's own beliefs, the mechanism merely has to copy representations from the Belief Box, embed the copies in a representation schema of the form *I believe that* ___, and then place the new representations back into the Belief Box. The proposed mechanism would work in much the same way to produce representations of one's own desires, intentions, and imaginings. (2003, 160–61).

One major lacuna in this account is its silence about an entire class of mental states: bodily feelings. They do not fit the model because, at least in the orthodox approach, sensations lack representational content, which is what the Nichols-Stich account relies upon. Their account is a syntactic theory, which says that the monitoring mechanism operates on the syntax of the mental representations monitored. A more general problem is what is meant by saying that the proposed mechanism would work in "much the same way" for attitude types other than belief. How does the proposed mechanism decide *which* attitude to ascribe? Which attitude verb should be inserted into the schema *I ATTITUDE that* ____? Should it be belief, desire, hope, fear, etc.? Each contentful mental state consists, at a minimum, in an attitude type plus a content. The Nichols-Stich theory deals only with contents, not types. In apparent recognition of the problem, Nichols and Stich make a parenthetical suggestion: perhaps a distinct but parallel mechanism exists for each attitude type. But what a profusion of mechanisms this would posit, with each mechanism essentially "duplicating" the others! Where is nature's parsimony that they appeal to elsewhere in their book?

The Nichols-Stich model of monitoring belongs to a family of self-attribution models that can be called "redeployment" theories because they try to explain (p. 419) self-attribution in terms of redeploying the content of a first-level mental state at a meta-representational level. Another such theory is that of Evans (1982), defended more recently by Gordon (1996), who calls it the "ascent-routine" theory. Gordon describes the ascent routine as follows: the way in which one determines whether or not one believes that *p* is simply to ask oneself the question whether or not *p*. The procedure is presumably to be completed as follows: if one answers the whether-*p* in the affirmative, one then "ascends" a level and also gives an affirmative answer to the question, "Do I think/believe that *p*?"

The ascent-routine theory faces a problem previously encountered with the monitoring theory. The basic procedure is described only for belief and lacks a clear parallel for classifying other attitudes or sensations. How is it supposed to work with hope, for example? Another problem concerns the procedure's details. When it says that a mindreader "answers" the whether-*p* question, what exactly does this mean? It cannot mean *vocalizing* an affirmative answer, because this will not cover cases of self-ascription where the answer is only *thought,* not vocalized. What apparently is meant by saying that one gives the "answer" *p* is that one *judges* the answer to be *p*. But how is one supposed to *tell* whether or not one judges that *p?* Is this not the same question of how one determines whether one (occurrently) believes that *p?* This is the same problem we started with, so no progress appears to have been made.

As we return to an introspectivist approach, notice that it is uncommitted to any strong view about introspection's reliability. Traditionally, introspection was associated with infallibility, but this is an easily detachable feature that few current proponents espouse. Introspectionism is often associated with a perceptual or quasi-perceptual model of self-knowledge, as the phrase "inner sense" suggests. Is that a viable direction? Shoemaker (1996) argues to the contrary. There are many disanalogies between outer sense and introspection, though not all of these should deter a theorist, says Shoemaker. Unlike

standard perceptual modalities, inner sense has no proprietary phenomenology, but this should not disqualify a quasi-perceptual analogy. A more serious disanalogy, according to Shoemaker, is the absence of any organ that orients introspection toward its cognitive objects (current mental states) in the manner in which the eyes or nose can be oriented toward their objects. Shoemaker considers but rejects attention as a candidate organ of introspection.

This rejection is premature, however. A new psychological technique called “descriptive experience sampling” has been devised by Hurlburt (Hurlburt and Heavey, 2001) for studying introspection. Subjects are cued at random times by a beeper, and they are supposed to pay immediate attention to their ongoing experience upon hearing the beep. This technique revealed thoughts of which they had not initially been aware, though they were not unconscious. Schooler and colleagues (2004) have made similar findings, indicating that attention is typically required to trigger reflective awareness via introspection. Actually, the term *introspection* is systematically ambiguous. It can refer to a process of inquiry, that is, inwardly directed attention, that chooses a selected state for analysis. Or it can refer to the process of performing an analysis of the state and outputting some description or classification of it. In the (p. 420) first sense, introspection itself is a form of attention, not something that requires attention in order to do its job. In the latter sense, it is an operation that performs an analysis or description of a state once attention has picked out the state to be analyzed or described.

If introspection is a perception-like operation, should it not include a transduction process? If so, this raises two questions: what are the inputs to the transduction process, and what are the outputs? Goldman (2006, 246–55) addresses these questions and proposes some answers. There has not yet been time for these proposals to receive critical attention, so it remains to be seen how this new quasi-perceptual account of introspection will be received. In any case, the problem of first-person mentalizing is as difficult and challenging as the problem of third-person mentalizing, though it has thus far received a much smaller dollop of attention, especially among cognitive scientists.

8. Conclusion

Like most topics at the cutting edge of either philosophy or cognitive science, mindreading is awash with competing theories and rival bodies of evidence. The landscape is especially difficult to negotiate because it involves investigations using a myriad of disparate methodologies, ranging from a priori reflection to the latest techniques of contemporary neuroscience. The resulting variety of evidential sources ensures that new and fascinating findings are always around the corner; but it also makes it likely that we will not see a settled resolution of the debate in the very near future. It would be misguided to conclude that the amount of research effort devoted to the subject is disproportionate to its importance. To the contrary, the target phenomenon is a key to human life and sociality. People's preoccupation with mindreading, arguably at multiple levels, is a fundamental facet of human nature, and a philosophico-scientific understanding of how we go about this task must rank as a pre-eminent intellectual desideratum for philosophy of mind and for cognitive science.

References

Adolphs, R., Tranel, D., and Damasio, A. R. (2003). Dissociable neural systems for recognizing emotions. *Brain and Cognition* 52: 61–69.

Baron-Cohen, S., Leslie, A., and Frith, U. (1985). Does the autistic child have a “theory of mind”? *Cognition* 21: 37–46.

—— . (1986). Mechanical, behavioral, and intentional understanding of picture stories in autistic children. *British Journal of Developmental Psychology* 4: 113–25.

(p. 421) Bavelas, J. B., Black, A., Lemery, C. R., and Mullett, J. (1986). “I show how you feel”: Motor mimicry as a communicative act. *Journal of Personality and Social Psychology* 50: 322–29.

Birch, S. A. J., and Bloom, P. (2003). Children are cursed: An asymmetric bias in mental-state attribution. *Psychological Science* 14: 283-86.

Buccino, G; Binkofski, F, Fink, G. R, Fadiga, L, Fogassi, L, Gallese, V, Seitz, R. J, Zilles, K, Rizzolatti, G, and Freund, H.-J. (2001). Action observation activates premotor and parietal areas in a somatotopic manner: An fMRI study. *European Journal of Neuroscience* 13(2): 400–404.

Buckner, R. L., and Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences* 11: 49–57.

Calder, A. J., Keane, J., Manes, F., Antoun, N., and Young, A. W. (2000). Impaired recognition and experience of disgust following brain injury. *Nature Reviews Neuroscience* 3: 1077–78.

Camerer, C., Loewenstein, G., and Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy* 97: 1232–54.

Carlson, S. M., and Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development* 72: 1032–53.

Carruthers, P. (2006). *The Architecture of the Mind.* Oxford: Oxford University Press.

Chartrand, T. L., and Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology* 76: 893–910.

Churchland, P. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy* 78: 67–90.

Collingwood, R. G. (1946). *The Idea of History*. Oxford: Clarendon Press.

Currie, G., and Ravenscroft,I. (2002). *Recreative Minds.* Oxford: Oxford University Press.

Damasio, A. R., Tranel, D., and Damasio, H. (1990) Face agnosia and the neural substrates of memory. *Annual Review of Neuroscience* 13: 89–109.

Dennett, D. C. (1987). *The Intentional Stance*. Cambridge, MA: MIT Press.

Dimberg, U., Thunberg, M., and Elmehed, K. (2000). Unconscious facial reactions to emotional facial expressions. *Psychological Science* 11: 86–88.

Evans, G. (1982). *The Varieties of Reference,* edited by J. McDowell. Oxford: Oxford University Press.

Fadiga, L., Fogassi, L, Pavesi, G., and Rizzolatti, G. (1995). Motor facilitation during action observation: A magnetic stimulation study. *Journal of Neurophysiology* 73: 2608–11.

Fodor, J. A. (1983). *The Modularity of Mind.* Cambridge, MA: MIT Press.

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

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., and Rizzolatti, G. (2005). Parietal lobe: From action organization to intention understanding. *Science* 308: 662–67.

Gallese, V., and Goldman, A. (1998). Mirror neurons and the simulation theory of mindreading. *Trends in Cognitive Sciences* 2: 493–501.

Gallese, V., Fadiga, L., Fogassi, L., and Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain* 119: 593–609.

Gergely, G., and Csibra, G. (2003). Teleological reasoning in infancy: The naïve theory of rational action. *Trends in Cognitive Sciences* 7: 287–92.

Gergely, G., Nadasdy, Z., Csibra, G., and Biro, S. (1995). Taking the intentional stance at 12 months of age. *Cognition* 56: 165–93.

Goldman, A. I. (1989). Interpretation psychologized. *Mind and Language* 4: 161–85.

—— . (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading.* New York: Oxford University Press.

(p. 422) —— . (2008). Mirroring, Mindreading, and Simulation. In J. Pineda (ed.), *Mirror Neuron Systems*: *The Role of Mirroring Processes in Social Cognition.* New York: Humana Press.

Goldman, A. I., and Sripada, C. S. (2005). Simulationist models of face-based emotion recognition. *Cognition* 94: 193–213.

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences* 16: 1–14.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., and Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review* 111: 3–32.

Gopnik, A., and Meltzoff, A. N. (1997). *Words, Thoughts and Theories*. Cambridge, MA: MIT Press.

Gopnik, A., and Wellman, H. (1992). Why the child's theory of mind really *is* a theory. *Mind and Language* 7: 145–71.

Gordon, R. M. (1986). Folk psychology as simulation. *Mind and Language* 1: 158–71.

—— . (1996).“Radical” Simulationism. In P. Carruthers and P. Smith (eds.), *Theories of Theories of Mind*. Cambridge: Cambridge University Press.

Hatfield, E., Cacioppo, J. T., and Rapson, R. L. (1994). *Emotional Contagion.* Cambridge: Cambridge University Press.

Heal, J. (1986). Replication and Functionalism. In J. Butterfield (ed.), *Language, Mind, and Logic*. Cambridge: Cambridge University Press.

Hurlburt, R. T., and Heavey, C. L. (2001). Telling what we know: Describing inner experience. *Trends in Cognitive Sciences* 5: 400–403.

Iacoboni, M., Molnar-Szakacs,I., Gallese, V., Buccino, G., Mazziotta, J. C., and Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS Biology* 3: 529–35.

Jackson, P. L., Meltzoff, A. N., and Decety, J. (2004). How do we perceive the pain of others? A window into the neural processes involved in empathy. *NeuroImage* 24: 771–79.

Jeannerod, M. (2001). Neural simulation of action: A unifying mechanism for motor cognition. *NeuroImage* 14: S103–9.

Johnson, S. C. (2005). Reasoning about intentionality in preverbal infants. In P. Carruthers, S. Laurence, and S. Stich (eds.), *The Innate Mind: Structure and Contents*. Oxford: Oxford University Press.

Johnson, S. C., Slaughter, V., and Carey, S. (1998). Whose gaze will infants follow? Features that elicit gaze-following in 12-month-olds. *Developmental Science* 1: 233–38.

Kanwisher, N., McDermott, J., and Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience* 17: 4302–11.

Keysers, C., Wicker, B., Gazzola, V., Anton, J-L., Fogassi, L., and Gallese, V. (2004). A touching sight: SII/PV activation during the observation of touch. *Neuron* 42: 335–46.

Kohler, E., Keysers, C., Umilta, M. A., Fogassi, L., Gallese, V., and Rizzolatti, G. (2002). Hearing sounds, understanding actions: Action representation in mirror neurons. *Science* 297: 846-48.

Kosslyn, S. M. (1978). Measuring the visual angle of the mind's eye. *Cognitive Psychology* 7: 341–70.

—— . (1994). Image and Brain: The Resolution of the Imagery Debate. Cambridge, MA: MIT Press.

Leslie, A. M. (1994). *Pretending* and *believing*: Issues in the theory of ToMM. *Cognition* 50: 211–38.

(p. 423) Leslie, A. M., and German, T. (1995). Knowledge and Ability in “Theory of Mind”: One-Eyed Overview of a Debate. In M. Davies and T. Stone (eds.), *Mental Simulation.* Oxford: Blackwell.

Lhermitte, F., Pillon, B., and Serdaru, M. (1986). Human autonomy and the frontal lobes. Part I: Imitation and utilization behavior: a neuropsychological study of 75 patients. *Annals of Neurology* 19: 326–34.

Meltzoff, A. N., and Moore, M. K. (1983). Newborn infants imitate adult facial gestures. *Child Development* 54: 702–9.

Mitchell, P., and Lacohee, H. (1991). Children's early understanding of false belief. *Cognition* 39: 107–27.

Morrison,I., Lloyd, D., de Pelligrino, G., and Roberts, N. (2004). Vicarious responses to pain in anterior cingulate cortex. Is empathy a multisensory issue? *Cognitive Affective Behavioral Neuroscience* 4: 270–78.

Nichols, S., and Stich, S. P. (2003). *Mindreading*. Oxford: Oxford University Press.

Nisbett, R., and Wilson, T. (1977). Telling more than we can know. *Psychological Review* 84: 231–59.

Onishi, K. H., and Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science* 308: 255–58.

Paccalin, C., and Jeannerod, M. (2000). Changes in breathing during observation of effortful actions. *Brain Research* 862: 194–200,

Perner, J. (1991). *Understanding the Representational Mind*. Cambridge, MA: MIT Press.

De Renzi, E., Cavalleri, F., and Facchini, S. (1996). Imitation and utilization behavior. *Journal of Neurology and Neurosurgical Psychiatry* 61: 396–400.

Rizzolatti, G., Fadiga, L., Gallese, V., and Foggasi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research* 3: 31–41.

Samson, D., Apperly,I. A., Kathirgamanathan, U., and Humphreys, G. W. (2005). Seeing it my way: A case of a selective deficit in inhibiting self-perspective. *Brain* 128: 1102–11.

Saxe, R. (2005). Against simulation: The argument from error. *Trends in Cognitive Sciences* 9: 174–79.

Scholl, B., and Leslie, A. M. (1999). Modularity, development and “theory of mind”. *Mind and Language* 14: 131–53.

—— . (2001). Minds, modules and meta-analysis. *Child Development* 72: 696–701.

Schooler, J., Reichle, E. D., and Halpern, D. V. (2004). Zoning-out during reading: Evidence for dissociations between experience and meta-consciousness. In D. Levin (ed.), *Thinking and Seeing: Visual Meta-Cognition in Adults and Children*. Cambridge, MA: MIT Press.

Schulz, L. E., and Gopnik, A. (2004). Causal learning across domains. *Developmental Psychology* 40: 162–76.

Sellars, W. (1956). Empiricism and the Philosophy of Mind. In H. Feigl and M. Scriven (eds.), *Minnesota Studies in Philosophy of Science*, vol. 1. Minneapolis: University of Minnesota Press.

Shoemaker, S. (1996). *The First-Person Perspective and Other Essays*. New York: Cambridge University Press.

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R., and Frith, C. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science* 303: 1157–62.

Spelke, E. (1994). Initial knowledge: Six suggestions. *Cognition* 50: 431–45.

Stich, Stephen (1981). Dennett on intentional systems. *Philosophical Topics* 12: 38–62.

Stueber, K. (2006). *Rediscovering Empathy*. Cambridge, MA: MIT Press.

(p. 424) Tager-Flusberg, H. (2000). Language and Understanding Minds: Connections in Autism. In S. Baron-Cohen, H. Tager-Flusberg, and D. Cohen (eds.), *Understanding Other Minds: Perspectives from Developmental Cognitive Neuroscience*, 2nd ed. Oxford: Oxford University Press.

Van Boven, L., and Loewenstein, G. (2003). Social projection of transient drive states. *Personality and Social Psychology Bulletin* 29: 1159–68.

Van Boven, L., Dunning, D., and Loewenstein, G. (2000). Egocentric empathy gaps between owners and buyers: Misperceptions of the endowment effect. *Journal of Personality and Social Psychology* 79: 66–76.

Wellman, H. M., Cross, D., and Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development* 72: 655–84.

Wicker, B., Keysers, C., Plailly, J., Royet, J-P., Gallese, V., and Rizzolatti, G. (2003). Both of us disgusted in *my* insula: The common neural basis of seeing and feeling disgust. *Neuron* 40: 655–64.

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

Wimmer, H., and Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* 13: 103–28.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition* 69: 1–34.

Zaitchik, D. (1991). Is only seeing really believing? Sources of the true belief in the false belief task. *Cognitive Development* 6: 91–103.

Alvin I. Goldman

Alvin I. Goldman, Department of Philosophy and Center for Cognitive Science, Rutgers University