UNIVERSITY OF CALIFORNIA, SAN DIEGO

Beyond Folk Psychology? Toward an Enriched Account of Social Understanding

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Philosophy and Cognitive Science

by

Mitchell Albert Herschbach

Committee in charge:

Professor William Bechtel, Chair
Professor Paul Churchland
Professor Gedeon Deák
Professor Rick Grush
Professor Rafael Núñez

2010

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

The Dissertation of Mitchell Albert Herschbach is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2010

TABLE OF CONTENTS

ACKNOWLEDGMENTS

I have many people to thank for their help in conceiving this project and bringing it to completion. I am very grateful to Bill Bechtel for hiring me as Editorial Assistant for *Philosophical Psychology*, which eventually led to his serving as my dissertation advisor. Bill has been an incredible source of encouragement and wisdom since the very initial stages of the project. I have also learned a tremendous amount from Paul Churchland and Rick Grush. Along with Bill, they have served as exceptional models for how to do empirically informed research in philosophy. On the cognitive science side, I have benefited greatly from courses and conversations with Gedeon Deák and Rafael Núñez.

I would also like to thank my family and friends. My fellow graduate students have been invaluable as colleagues and friends. My friends outside the academic world have been just as important in encouraging my intellectual endeavors while keeping me grounded. My parents, John and Linda, have supported me in everything I have done, including doctoral studies. I would not be the person I am today without them. Finally, I would like to thank Alexis Rochlin. I am so thankful for her love and support during every stage of this project, and for helping to remind me of what is most important in life.

Chapters 1 and 5 contain material reproduced from "False-belief understanding and the phenomenological critics of folk psychology," *Journal of Consciousness Studies*, *15*(12), 33–56. Permission to reproduce this material has been granted by the copyright owner, Imprint Academic. The dissertation author was the primary investigator and author of this paper.

Chapter 4 is a reprint, with slight modifications, of "Folk psychological and phenomenological accounts of social perception," *Philosophical Explorations*, *11*(3), 223–235. Permission to reproduce this material has been granted by the copyright owner, Taylor & Francis. The dissertation author was the primary investigator and author of this paper.

CURRICULUM VITA

Education

2010 University of California, San Diego
Ph.D. in Philosophy and Cognitive Science

2007 University of California, San Diego
M.A. in Philosophy

2002 Santa Clara University
B.A. in Philosophy, B.S. in Psychology, *summa cum laude*

Publications

Journal Articles

Herschbach, M. (2008a). False-belief understanding and the phenomenological critics of folk psychology. *Journal of Consciousness Studies*, *15*(12), 33–56.

Herschbach, M. (2008b). Folk psychological and phenomenological accounts of social perception. *Philosophical Explorations*, *11*(3), 223–235.

Book Chapters

Bechtel, W., & Herschbach, M. (2010). Philosophy of the cognitive sciences. In Fritz Allhoff (Ed.), *Philosophies of the Sciences* (pp. 239–261). Oxford: Wiley-Blackwell.

Conference Proceedings

Herschbach, M. (2008). The concept of simulation in control-theoretic accounts of motor control and action perception. In V. Sloutsky, B. Love, and K. McRae (Eds.), *Proceedings of the 30th annual conference of the Cognitive Science Society* (pp. 315–320). Austin, TX: Cognitive Science Society.

ABSTRACT OF THE DISSERTATION

Beyond Folk Psychology? Toward an Enriched Account of Social Understanding

by

Mitchell Albert Herschbach

Doctor of Philosophy in Philosophy and Cognitive Science

University of California, San Diego, 2010

Professor William Bechtel, Chair

Folk psychology is the ability to interpret people's mental states (beliefs, desires, etc.) and use this information to explain and predict their behavior. While folk psychology has traditionally been seen as fundamental to human social understanding, philosophers drawing on the phenomenological tradition have recently argued that most of our everyday social interactions do not involve folk psychology. I defend the role of folk psychology in human social understanding against these

phenomenological critics. I argue that we need not abandon the folk psychological picture to heed the central claims of these phenomenological critics. In so doing, I develop an enriched account of human social understanding that accepts their descriptions of the phenomena of human social understanding while retaining a significant role for folk psychological reasoning at the subpersonal level.

In chapter 1 I describe the traditional folk psychological account of social understanding and the challenge to it raised by these critics. Since it assumes folk psychology is pervasive, the traditional philosophical and empirical research on human social understanding focuses on the psychological processes by which we attribute mental states: whether we apply theoretical knowledge about human psychology, as proposed by the theory theory, and/or use own psychological mechanisms to "simulate" other people's mental states, as the simulation theory suggests. Two central claims made by the phenomenological critics against this traditional picture are: (i) that some mental state understanding occurs by "direct perception," without the need for theorizing or simulation; and (ii) that attributing beliefs and desires is not required and not often used for unreflectively interacting with other people. I argue that direct social perception and unreflective social interaction are phenomena that should be better emphasized in accounts of human social understanding, but which can be explained by folk psychological reasoning occurring at the subpersonal level, outside of conscious awareness. In chapters 2–3 I develop my conception of personal and subpersonal levels. I then apply this framework in the next two chapters to argue that direct social perception (chapter 4) and unreflective social

interaction (chapter 5) can, contrary to the phenomenological critics, be driven by folk psychological theorizing and/or simulation.

Introduction

A recent trend in philosophy of mind and the cognitive sciences has been to draw on resources from the philosophical tradition of phenomenology, whose major figures include Edmund Husserl, Martin Heidegger, and Maurice Merleau-Ponty. While Hubert Dreyfus (e.g., 1972, 1991) has long championed the relevance of phenomenology (particularly Heidegger) to the cognitive sciences, the last decade or so has seen a huge spike in this interdisciplinary endeavor. This can be seen in the publication of a number of edited volumes (e.g., Petitot, Varela, Pachoud, & Roy, 1999; Smith & Thomasson, 2005), the journal *Phenomenology and the Cognitive Sciences*, and in particular Shaun Gallagher and Dan Zahavi's *The Phenomenological Mind* (2008), which has been described as "the first systematic overview of philosophy of mind from a phenomenological angle" (Slors, 2008, p. 34). The phenomenological perspective has been argued to be relevant to a number of issues in the cognitive sciences, including the nature of intentionality, consciousness and self-consciousness, the experience of time, embodiment, and action.

In this dissertation I will focus on what phenomenologists have to say about the nature of human social understanding—particularly our understanding of other people's minds, which has often been called "folk psychology." As I describe in chapter 1, the standard view in philosophy of mind and the cognitive sciences is that folk psychology plays a pervasive role in our navigation of the social world. Assuming this, much of the research on human social understanding has explored the psychological processes by which we attribute mental states to other people: whether

we make use of theoretical knowledge about human psychology, as the *theory theory* (TT) proposes; "simulate" other people's mental states, as the *simulation theory* (ST) claims; or use some combination of the two. Philosophers working in the phenomenological tradition, in particular Shaun Gallagher, Dan Zahavi, and Matthew Ratcliffe, have called into question this folk psychological picture, arguing that it vastly overestimates the importance of folk psychology in our daily lives and thus mischaracterizes the nature of human social understanding. While admitting that we do sometimes consciously theorize about or simulate other people's mental states, they argue that most of the time we understand and interact with people through other means. According to these phenomenological critics, accounts of human social understanding should better attend to our experience of other people when describing the phenomena constituting human social understanding, and consider alternatives to TT and ST when explaining the psychological processes underlying this experience.

It could be said, therefore, that these phenomenological critics think accounts of human social understanding must move beyond folk psychology. I will argue that this position is too radical. I articulate instead a synthesis of the folk psychological and phenomenological accounts, accepting some of the phenomenological critics' main points without abandoning the importance of folk psychology to our daily lives. I endorse and expand upon the phenomenologists' call to enrich accounts of human social understanding by, for instance, (a) including phenomenologically-direct social perception as a mode of access to others' mental states distinct from conscious theorizing or simulation, and (b) emphasizing unreflective social interaction (in addition to the reflective social judgments studied in standard psychology

experiments) and nonmentalistic modes of social understanding. I argue, however, that this can be done while retaining a significant role for folk psychology, and for TT and ST as potential explanations of our folk psychological abilities.

Central to my analysis is distinguishing different levels of phenomena and explanation in the cognitive sciences—in particular, the distinction between the *personal level* of whole organisms and their experiences, and *subpersonal levels* concerning the parts of organisms, especially their brains. So after articulating the debate between folk psychological and phenomenological accounts of human social understanding in chapter 1, in chapter 2 I develop my account of personal and subpersonal levels. I start by explicating Daniel Dennett's (1969) original presentation of personal and subpersonal levels, and how his view changes in subsequent writings. I then sketch my mechanistic account of the personal–subpersonal distinction, where subpersonal-level mechanisms are offered as explanations of personal-level phenomena. Through the work of John McDowell I characterize how I fit phenomenological claims about conscious experience within this mechanistic framework, and through the work of Susan Hurley discuss how this framework can accommodate claims about the environmentally situated nature of cognition. I conclude chapter 2 by showing the compatibility between my account of levels and that of the phenomenological critics.

In chapter 3 I focus on the methods of inquiry used to obtain multilevel accounts of cognitive phenomena. I first describe third-person research techniques found in the cognitive sciences, such as behavioral experiments and methods for directly investigating the structure and activity of our neural mechanisms. I then

address first-person methods used to acquire more direct evidence about the structure and contents of consciousness. Here I describe methods derived from the phenomenological tradition, and how they might be integrated with third-person data. I use this explanatory and investigative framework in chapters 4–5 to examine specific personal- and subpersonal-level features of social understanding addressed by the phenomenological critics. I argue that phenomenological critics have offered no reason to deny that the folk psychological picture can accommodate these phenomena, as long as mental state attribution and the theorizing and/or simulation driving these attributions are treated as *subpersonal-level* phenomena.

In chapter 4 I tackle the phenomenologists' claim that we can "directly perceive" some mental states, such as emotions and goals or intentions. Phenomenologists see these phenomena as challenging folk psychological accounts, which require a psychological step beyond perception to infer other's "unobservable" mental states from their "observable" behavior. I argue that the direct perception of mental states is indeed in conflict with personal-level versions of TT and ST, but not with subpersonal-level versions of each. Focusing specifically on Dan Zahavi's arguments against TT and Shaun Gallagher's arguments against ST, I argue that neither gives us reason to reject subpersonal-level versions of TT and ST. The folk psychological picture is in this way able to accommodate a class of phenomena identified by the phenomenological critics.

In chapter 5, I move beyond emotion and goal/intention attribution to full-blown folk psychology involving the attribution of beliefs and desires. The phenomenological critics argue that belief–desire attribution is restricted to conscious,

reflective cognition, and is not used for unreflective social interaction. Having already defended subpersonal-level versions of TT and ST, this leaves open the possibility that unreflective social interaction is driven by subpersonal-level attribution of beliefs and desires. But what is the empirical evidence for this view? As the phenomenologists note, the standard experimental research does not speak directly to the role of folk psychology in unreflective social interaction. I argue, however, that more recent experimental research on our understanding of other people's false beliefs—widely recognized as a paradigmatic case of folk psychological understanding—provides direct evidence of folk psychology driving unreflective social interaction. I then defend my interpretation of these experiments against possible objections from the phenomenological critics.

What emerges from this engagement with the phenomenological critics is an enriched account of human social understanding, including phenomena not often mentioned in traditional folk psychological accounts, such as direct social perception and nonmentalistic forms of social understanding, and ones not directly addressed in the traditional empirical research, such as the online use of folk psychology.

Chapter 1. The Phenomenological Critics of Folk Psychology

1. Social Understanding and "Folk Psychology"

Over the last few decades, philosophical and empirical discussions of human social understanding have been formulated in terms of "folk" or "commonsense" psychology: that it is our folk understanding of human psychology which explains our abilities to understand and interact with other people (e.g., Carruthers & Smith, 1996; Davies & Stone, 1995a, 1995b; Goldman, 2006; Nichols & Stich, 2003). Talk of folk psychology[1] is so commonplace now that it may seem to be just a catchall term for whatever knowledge of human psychology and behavior most humans possess that makes possible human social interaction.[2] But folk psychology as usually

[1] "Theory of mind" is another phrase often used in place of "folk psychology." I use the latter term because "theory of mind" too easily brings to mind the "theory theory" account of our folk psychological abilities. Admittedly, "folk psychology" has the unfortunate implication that it is "folksy" and somehow unsophisticated. But I will use it rather than other alternatives in the literature, such as "mindreading" or "mentalizing," as a term for the characterization of human social understanding common to theory theory and simulation theory. But as I will show below, it is possible that the folk psychological account of social understanding is mistaken.

[2] That "all social understanding [is] a matter of the attribution of mental states and the deployment of those attributed states to explain and predict behavior" is what Bermúdez (2003, 2005) terms the "broad construal" of the domain of folk psychology. The "narrow construal" restricts folk psychology to only "those occasions on which we explicitly and consciously deploy the concepts of folk psychology in the services of explanation and/or prediction" (Bermúdez, 2005, p. 176). Bermúdez's distinction is different from Stich & Ravencroft's (1994) contrast between "internalist" and "externalist" senses of folk psychology. For Stich & Ravencroft, folk psychology can be conceived of as either: (1) the internally represented knowledge that is used in the description, explanation, and (verbal and nonverbal) prediction of behavior (the *internal* account of folk psychology); or (2) as the set of abstract psychology generalizations which most people would recognize and assent to, where these generalizations are not part of the internal mechanism supporting explanation and prediction of behavior (the *external* account of folk psychology). Bermúdez's narrow–broad distinction concerns whether folk-psychology-internal plays a role only in explicit, conscious explanation and prediction, or in all forms of social understanding. This is the issue with which I am concerned here.

characterized is not so theoretically uncommitted, instead involving specific assumptions about the nature of interpersonal understanding, and the role of such understanding in social interaction. As Ratcliffe and Hutto (2007) summarize:

> the received wisdom about folk psychology encapsulates two chief assumptions: (i) that making sense of actions requires interpreting them in terms of reasons composed of various propositional attitudes (at a bare minimum—beliefs and desires) and (ii) that this activity is primarily concerned with providing predictions and explanations of actions. (p. 2)

The two most dominant accounts of folk psychology, *theory theory* (TT) and *simulation theory* (ST),[3] share these basic assumptions, treating folk psychology as involving the attribution of propositional attitudes and other kinds of mental states to others in order to explain and predict their behavior. The core idea of TT is that mental state attribution requires the possession of a folk psychological theory, i.e., a set of generalizations about the relationships between environmental conditions, mental states, and behavior. Behavioral predictions, for example, may be made by using information about a person's mental states and the relevant folk psychological generalizations about the decision-making of people with those kinds of mental states, to infer what they will decide to do. The core idea of ST is that we do not need theoretical knowledge of folk psychology to understand other people, but instead use our own minds to "simulate" the other person's mental states and processes, exploiting the similarity between one's own psychological makeup and that of other people. In

[3] Rationality theory (RT), as developed in the work of Daniel Dennett and Donald Davidson, is a third contender in this debate. The essential feature of RT is the claim that when we attribute mental states to others, we assume others are rational agents. I agree with Goldman's (2006) assessment that "RT is no longer a serious rival to TT and ST" (p. 67), so I will not discuss it much in this dissertation.

the case of behavioral prediction, ST proposes that I "imagine" having the mental states of the other person and engage in "pretend" decision-making given those mental states, in order to predict their future mental states and behavior. Much ink has been spilled attempting to formulate and distinguish TT and ST, with many researchers moving towards hybrid theory–simulation accounts, taking both theory and simulation to have necessary roles. For most philosophers and scientists, TT, ST, or some combination of the two, are the only theoretical options and the assumptions they share go unquestioned. Folk psychology is presumed to be the proper way to frame any investigation of the nature of human social understanding, whether the behavioral experiments of developmental psychologists, neuroimaging studies of neuroscientists, or the theoretical work of philosophers. It is our proficiency with folk psychology that is offered as an explanation for human social interaction. As psychologist Helen Tager-Flusberg (2005) has recently put it:

> Successful social interactions depend on the ability to understand other people's behavior in terms of their mental states, such as beliefs, desires, knowledge, and intentions. Social situations and events cannot be interpreted on the basis of overt behavior without representing the mental states underlying people's actions. Understanding people as intentional, mental beings is at the core of social cognition, within which the ability to interpret people's behavior in a mentalistic explanatory framework using a coherent, causally related set of mental constructs is central to a theory of mind [or folk psychology]. (p. 276)

These assumptions of folk psychology can be made more concrete by examining a dominant experimental paradigm for studying social understanding: so-called "third-person false-belief tasks," which study the ability of a given individual to understand that other people can have false beliefs. A common form of these tasks involves inducing a false belief in someone about the location of an object by moving

it without their knowledge. In Baron-Cohen, Leslie, and Frith's (1985) version, often called the Sally–Anne task, dolls named Sally and Anne are used in place of actual humans as protagonists. The child-subject observes Sally put a marble in a basket then leave the room. While Sally is away, Anne enters and moves the marble from the basket into a box in the same room. Sally then reenters, and the child is asked, "Where will Sally look for the marble?" The child knows that the marble is currently in the box. The child must suppress this information, and consider where Sally thinks the marble is located in order to predict Sally's behavior. The child must remember that Sally last saw the marble in the basket and that she is unaware of Anne's having moved the marble. Thus the correct prediction is that Sally will look for the marble in the basket. If asked to explain this behavior, the child should say Sally will act this way because she believes the marble to be located there, even though it is actually in the box.

Passing false-belief tasks is taken as signaling a significant development in the acquisition of a folk psychological understanding of other people. The standard interpretation is that such understanding requires the child to possess a concept of BELIEF (along with other concepts such as DESIRE and INTENTION) and thereby understand the representational nature of belief, i.e., that people's beliefs can fail to accord with reality. Theory theorists propose that this developmental milestone is due to the acquisition of a more advanced folk psychological theory (whether through learning or maturation) that includes the mature concept of BELIEF. Simulation theorists, in contrast, usually argue that being able to pass false-belief tasks is not due to conceptual change, but rather is due to development in the ability to simulate other

people's beliefs, such as the ability to inhibit one's own beliefs about a particular subject matter in order to create and make use of a simulation of others' beliefs which diverge from one's own and/or from reality. Whatever the story about its development, having a representational understanding of mental states is crucial to folk psychology's conception of social understanding. If people's actions are driven by their representations of the world, rather than merely how the world actually is, this must be appreciated for the full range of human behavior to be understood. Having such understanding can expand the forms of social interaction available to an agent. As the Sally–Anne task illustrates, it can allow the prediction of others' behavior driven by false beliefs. Exploiting this understanding of behavior could allow the manipulation of other's representational mental states in order to deceive them, which can be helpful in competitive situations. While theory-theorists and simulation theorists disagree about what exactly explains the ability to pass the false-belief task, both take such tasks to be paradigmatic of our folk psychological social understanding, and success in this domain to be the result of mental state attributions.[4]

Although false-belief tasks are a well-used example, other so-called "theory of mind" tasks from developmental psychology involve the same conception of mentalistic social understanding—they require someone to explain or predict another

[4] There is disagreement amongst researchers about how much weight should be placed on passing the false-belief task. Some treat failure on standard false-belief tasks as signaling the lack of the concept of BELIEF, and thus a deficient "theory of mind" or folk psychology. Others treat the false-belief task as a relatively difficult task that does not necessarily provide evidence about the folk psychological knowledge of a subject, since other task demands may mask their conceptual knowledge (i.e., their possession of the concept BELIEF and thus the ability to appreciate the false beliefs of others). See Bloom and German (2000) for discussion of why the standard false-belief task is not the ultimate test of folk psychological knowledge many researchers have taken it to be.

person's behavior by understanding their mental causes.[5] Wellman and Liu's (2004) battery of theory of mind tasks, for example, tests (in addition to false-belief understanding) the understanding of diverse desires (that people can act on desires which differ from one's own), diverse beliefs (that people can act on beliefs which differ from one's own), knowledge access (that perception can be required for knowledge), and the relation between belief and emotion (e.g., reporting that Teddy, who likes Cheerios, will be happy when he receives a box of Cheerios, but sad when he finds out that the Cheerios box is filled with rocks). In these tests, children are read a scenario, which is accompanied with picture props depicting objects, situations, and facial expressions. Children are asked explicit questions requiring them to predict a protagonist's behavior (e.g., which of two objects a person will choose, to test their understanding of the person's desires), or make judgments about their mental states (e.g., judging that someone will not have knowledge of the unexpected contents of a drawer when they have not seen inside of it).

Neuroimaging studies investigating the neural mechanisms underlying social understanding use similar behavioral tasks. One common method is to scan participants while they read and answer questions about "theory of mind stories" (i.e., stories which required attributing mental states to explain the behavior of the protagonists) in contrast to "non-theory of mind stories" (i.e., stories involving

[5] Although it is not always emphasized in the empirical literature, folk psychology treats mental states not just as causing behavior, but as providing *reasons* for people's actions. Hutto (2008a) emphasizes that understanding beliefs and desires is distinct from understanding reasons, for one could attribute individual beliefs and desires to people without understanding how these interrelate so as to provide reasons for their behavior.

physical causation, which do not require mental state attributions to be understood) and two sets of unlinked sentences (Fletcher, Happé, Frith, & Baker, 1995; H. L. Gallagher et al., 2000; Vogeley et al., 2001). The following is an example of a "theory of mind story" used in these studies:

> A burglar who has just robbed a shop is making his getaway. As he is running home, a policeman on his beat sees him drop his glove. He doesn't know the man is a burglar, he just wants to tell him he dropped his glove. But when the policeman shouts out to the burglar, "Hey, you! Stop!", the burglar turns round, see the policeman and gives himself up. He puts his hands up and admits that he did the break-in at the local shop.

After reading this passage, subjects were asked, "Why did the burglar do this?" To explain the burglar's behavior, the subject would have to attribute to the burglar the (mistaken) belief that the policeman knew he robbed the shop and was trying to apprehend him. Nonverbal stimuli, such as captionless cartoons, have also been used to test mental state understanding (H. L. Gallagher et al., 2000). The brains areas differentially activated during the "theory of mind" tasks compared to the other tasks are thought to be the locus of the neural mechanisms underlying mental state attribution.

2. The Phenomenological Critics of Folk Psychology

The folk psychological picture of human social understanding has come under attack recently from a number of directions, calling into question the core assumptions of the folk psychological account: that our standard way of understanding other people involves interpreting their mental states, and explaining or predicting behavior in

terms of these mental states. One prominent line of attack has come from philosophers working in the phenomenological tradition, such as Shaun Gallagher, Matthew Ratcliffe, and Dan Zahavi (Gallagher, 2001, 2004, 2005, 2007; Gallagher & Zahavi, 2008; Gallagher & Hutto, 2008; Ratcliffe, 2006a, 2006b, 2007; Zahavi, 2001, 2004a, 2005, 2007). Others working more from the analytic tradition, such as Jose Luis Bermúdez (2003, 2005) and Dan Hutto (2004, 2007, 2008a, 2008b), have made similar arguments, but I will focus here on the arguments of the "phenomenological critics," as I'll call them. While (usually[6]) acknowledging that some instances of social understanding involve folk psychological theorizing or simulation, the phenomenological critics argue that the folk psychological account fails to capture most cases of everyday social understanding. As Shaun Gallagher (2001) puts it: "Theory theory and simulation theory, at best, explain a very narrow and specialized set of cognitive processes that we sometimes use to relate to others.… Neither theoretical nor simulation strategies constitute the primary way in which we relate to, interact with or understand others" (p. 85).

Unsurprisingly, one major source of evidence appealed to by these phenomenological critics of folk psychology is the *phenomenology* of social

[6] While the phenomenological critics often accept that TT and ST are true but narrow theories of human social understanding, some of them have argued that these theories indeed fail to account for any of the phenomena of human social understanding. Matthew Ratcliffe (2007) pushes such a line that the folk psychology found in the philosophical and empirical literature fails to capture the nature of our everyday social understanding, both where it involves understanding of others' mental states and when it does not. While Ratcliffe admits that we do sometimes attribute mental states in the course of understanding other people's behavior, he thinks the folk psychological account inaccurately describes the nature of these phenomena—i.e., that folk psychology as commonly understood fails to exist. In this dissertation will not be evaluating as a whole Ratcliffe's "eliminativist" thesis about folk psychology, but where relevant will address specific criticisms of the folk psychological picture raised by Ratcliffe.

understanding, i.e., what our conscious experience of other people is like—where this includes both fine-grained descriptions of what we do find in social experience, and sometimes more importantly, what is *not* part of our experience. They appeal to such phenomenological descriptions offered by figures in the phenomenological tradition such as Edmund Husserl (e.g., 1950/1999), Martin Heidegger (e.g., 1927/1962), Maurice Merleau-Ponty (e.g., 1945/2002), as well their own examples. But descriptions of our experience of other people are not all the phenomenological critics have to offer. Their arguments often draw upon empirical research from the cognitive sciences, as well as the standard philosophical method of conceptual analysis. Thus, the phenomenological critique of folk psychological cannot be dismissed simply by rejecting phenomenological methods. Even if we dispute these methods, we should take seriously their descriptions of the phenomena of human social understanding, as well as the other kinds of evidence and argumentation they bring to bear on this issue. While I will be rejecting aspects of the phenomenologists' critique of folk psychology, I will argue that some of their insights can be synthesized with the traditional folk psychological accounts, TT and ST, to produce an enriched and more accurate account of human social understanding.

First, more needs to be said to introduce the phenomenologists' critique of folk psychology. These phenomenological critics call into question how we describe and delineate the phenomena of human social understanding, as well as the popular explanations of the psychological processes driving these phenomena, i.e., TT and ST. Their focus is on the class of phenomena I'll refer to as *unreflective social understanding,* which can be characterized negatively as cases of social understanding

where we do not consciously or explicitly attribute mental states by processes of theorizing or simulation. The general thesis of the phenomenological critics can thus be characterized as the claim that *the folk psychological picture fails to capture unreflective social understanding*. The scope of their critique is thus quite large, making it difficult to systematically analyze everything they have to say about the nature of human social understanding. We can, however, hone in on two central aspects of their critique by distinguishing two aspects of the folk psychological picture: (i) the general idea that social understanding involves attributing mental states, and (ii) that mental state attribution is produced by psychological processes of theorizing, as suggested by TT, or of simulation, as suggested by ST. I'll approach these in reverse order. According to the phenomenological critics, some phenomena of unreflective social understanding do involve mental state attribution, but are not well explained in terms of theorizing or simulation—i.e., with regard to the two aspects of the folk psychological picture indicated above, these phenomena are well described by (i) but not (ii). Here we find what the phenomenological critics describe as *social perception*: cases where we can *directly experience or perceive* certain mental states of other people in their expressive behavior, without the need for theorizing or simulation.[7] While they do not claim we can directly experience all types of mental

[7] Unfortunately, the phenomenological critics do not always admit that such phenomena actually involve mental state attribution. For example, Gallagher (e.g., 2001, 2005) tries to draw a distinction between "nonmentalistic" social perception and "mind-reading." The former is characterized as directly experiencing the "meaning" of people's actions and expressive movements, while the latter involves making inferences about "a hidden set of mental states (beliefs, desires, etc.)" (2001, p. 90). But "understanding the 'meaning' of people's actions" is just Gallagher's way of saying that we can immediately understand their intentions and emotions—which most researchers would call mental

states, they do believe we can directly perceive people's emotions and intentions without theorizing or simulation.

A second, more radical aspect of the phenomenological critics' agenda involves denying both features of the folk psychological picture: they argue that some aspects of our unreflective social understanding do not at all involve mental state attribution. To be more precise, the phenomenological critics are most concerned about the attributions of belief and desire which take pride of place in folk psychological accounts. Their claim is that belief–desire psychology (i.e., the understanding of other people's actions in terms of the beliefs and desires causing/rationalizing their behavior) is not required and not often used for unreflectively interacting with other people. Even if they allow for direct social perception (which I'm treating as a form of mental state attribution) in such cases, they

---

states. So why is Gallagher drawing this distinction between mentalistic and nonmentalistic modes of social understanding? It seems that Gallagher wants to restrict the term "mental state" to "hidden," "inner" states of people that we cannot directly observe. Because intentions and emotions are more intimately tied to bodily behavior, Gallagher wants to say they are not "mental" states in this sense of the term. More generally, it is a theme of the phenomenological approach to attack the Cartesian picture of mental states as conceptually distinct from bodily behavior (Zahavi, 2005). Phenomenologists try to carve a middle ground between Cartesian dualism (according to which others' minds are completely inaccessible to others) and behaviorism (according to which others' minds are completely accessible to others), where others' mental states "can be directly perceivable and yet retain a certain inaccessibility" (Overgaard, 2005, p. 249). But it is unclear exactly what states we usually call "mental" are claimed to be given this treatment. Emotions and intentions are used as the paradigm cases of directly perceivable states, but it is less clear if, say, beliefs and desires should also be treated in this way. Gallagher's distinction between mentalistic and nonmentalistic social understanding suggests the phenomenological account of the mental is not intended to apply to all states we normally call "mental." I am sympathetic to the idea that there are intentional states which are more intimately tied to our embodiment. But since there is not much clarity to this aspect of the phenomenological critics' account, I will continue with folk psychology's list of the states we call mental or psychological in nature, without committing myself to a particular account of the nature of these concepts. I will have a bit more to say about this issue in later chapters, especially chapters 4–5.

reject the idea that we must unreflectively attribute beliefs and desires to other people in order to interact with them.

In this dissertation, I will evaluate these two aspects of the phenomenologists' critique of folk psychology. While I will reject both facets of the phenomenologists' account, doing so will require enriching the folk psychological picture—with regard to the phenomena needing further attention in future research, and with regard to the conceptual resources used in explanations of these phenomena. I will now say a bit more about each of the two aspects of the phenomenologists' critique of folk psychology, and the arguments I'll be making about them in subsequent chapters.

2.1. Folk Psychology and Social Perception

The folk psychological account inherits the picture of social experience and mental state attribution found in traditional philosophical discussions of the metaphysics of mind and the epistemological problem of other minds. Given an understanding of the mind as "inner" and the body as "outer," it is assumed that while we possess direct access to our own minds, other people's minds are *inaccessible* or *unobservable* to us. Gilbert Ryle (1949/1984), for example, captures this traditional position as follows:

> …one person has no direct access of any sort to the events of the inner life of another. He cannot do better than make problematic inferences from the observed behavior of the other person's body to the states of mind which, by analogy from his own conduct, he supposes to be signalized by that behavior. Direct access to the workings of a mind is the privilege of that mind itself; in default of such privileged access, the workings of one mind are inevitably occult to everyone else. (p. 14).

The two dominant folk psychological accounts, TT and ST share this picture of mental states as unobservable, inner states of persons which can only be understood by others via some psychological step beyond perception. They differ in how they characterize what Ryle calls the "problematic inferences" from observed behavior to mental states (or, for example, from mental states to predicted future behavior). "Child scientist" versions of TT, for example, treat mental states as "abstract unobservable entities" postulated to account for the "evidential phenomena" to be explained, namely, "observable" behavior (e.g., Gopnik & Wellman, 1992). To "go beyond" the mere behaviors we find in experience, we must make theoretical inferences using a folk psychological theory. ST similarly claims we must use an additional psychological process to "go beyond" the perception of mere behavior and attribute mental states. But rather than making theoretical inferences, ST (at least Alvin Goldman's influential version of it) is closer to the view expressed by Ryle above that mental state attribution involves a sort analogical reasoning, where I attribute to the other the mental states and processes I would have if I were them.

Therefore, all folk psychological accounts share the idea that what we can *perceive*, what is "given" in experience, is only physical behavior. Simon Baron-Cohen's (1995) use of the term "mindblindness" for the lack of mental state understanding characteristic of people with autism is thus somewhat ironic, since no one in the folk psychological camp thinks we can literally "see" others' mental states. The phenomenological critics would instead say that the folk psychological account mistakenly characterizes *unimpaired* social perceivers as "mindblind" by treating our immediate experience as restricted to mere behavior. Against this folk psychological

picture, the phenomenological critics claim that we can indeed "directly perceive" some mental states of other people, such as their emotions and intentions. Following figures from the phenomenological tradition such as Husserl (1950/1999), Max Scheler (1954), and Edith Stein (1964), the phenomenological critics often use the term "empathy" (the standard translation of the German word "Einfühlung") to characterize this direct experience of the mental lives of other people (Thompson, 2001; Zahavi, 2005, ch. 6, 2007, 2008). But given the association in most circles with emotional empathy, I will mostly stick with the term "social perception." Scheler (1954), for example, uses such perceptual language in the following passage, often quoted by the phenomenological critics:

> For we certainly believe ourselves to be directly acquainted with another person's joy in his laughter, with his sorrow and pain in his tears, with his shame in his blushing, with his entreaty in his outstretched hands, with his love in his look of affection, with his rage in the gnashing of his teeth, with his threats in the clenching of his fist, and with the tenor of his thoughts in the sound of his words. If anyone tells me that this is not "perception," for it cannot be so, in view of the fact that a perception is simply a "complex of physical sensations," and that there is certainly no sensation of another person's mind nor any stimulus from such a source, I would beg him to turn aside from such questionable theories and address himself to the phenomenological facts. (p. 260)

In contrast to the folk psychological picture, Scheler and the recent phenomenological critics claim that our perceptual experience of other people is not of mere behavior, but can include as part of its content their emotions and intentions.

There is certainly a tension between folk psychological and phenomenological accounts of social perception, which I will attempt to adjudicate. The main issue is to determine whether TT and ST are inadequate as accounts of social perception, as the

critics claim. The step I see as most crucial to navigating this debate is to clearly distinguish the *phenomenology* of social perception from *explanations* describing the neural mechanisms enabling social perception. I will be explicating this contrast in terms of the distinction between *personal and subpersonal levels*. We certainly want to respect the phenomenological facts about social perception; this is something which I believe many philosophical discussions under the heading of folk psychology have failed to do adequately. But I contend that phenomenologists have been too quick in dismissing ST and TT. The arguments against simulation's playing a role in social perception (e.g., Gallagher, 2007) place too much weight on simulation's being a personal-level concept, and fail to recognize a role for the concept of simulation at the subpersonal level.[8] And while there are legitimate concerns with imposing the vocabulary of theoretical inference upon subpersonal mechanisms, there is certainly information processing of some sort being performed that goes beyond simulation, and which must be accounted for when explaining social perception.

Social perception is an explanandum not typically addressed as such in the folk psychological literature, largely because of a failure to recognize phenomenological facts about our social experience. The phenomenological critics have made progress in

[8] Greenwood (1999) argues that ST need not be restricted to experimental paradigms requiring subjects to verbally predict or explain behavior. He contends that even early formulations of ST referred to the nonverbal "anticipation" of behavior as being explained by unconscious simulation. As Greenwood notes, accounting for the relation between nonverbal anticipation and verbal predictions and explanations is an important task for ST. Greenwood calls into question whether ST should even be attempting to explain our verbal explanations of behavior. In the terminology I develop later in this dissertation, Greenwood recognizes the distinction between online and offline forms of social understanding, and suggests that ST may be better at accounting for online nonverbal anticipation of others' behavior, than for offline verbal explanation and prediction. This highlights that looking at online social understanding might require adjusting our explanations of offline social understanding.

describing this explanandum, but, I will argue, they have not done much in offering explanations of social perception, and their negative claims about TT and ST are not at all conclusive. This engagement with the critics, however, pushes defenders of the folk psychological account to be clearer about the level at which their views are being pitched—e.g., it would require them to explicitly defend TT and ST as subpersonal-level accounts of the psychological processes enabling social perception. As we'll see, this distinction between personal and subpersonal levels will also be crucial to my evaluation of the second aspect of the phenomenological critique of folk psychology, to which I will now turn.

2.2. Folk Psychology and Online Social Understanding

Phenomenological discussions of human social understanding have not been restricted to issues of social perception or empathy. Although not always recognized, phenomenologists have gone "beyond empathy" (Zahavi, 2005, p. 163) to investigate other ways people feature in our experience besides as targets of explicit mental state attributions. In the language of the phenomenologists, empathy involves "thematizing," making a focus of explicit awareness, the mental states of others. They challenge the idea that most of our interactions with other people involve this conscious, explicit experience of people's mental states, whether of the reflective theorizing or simulating emphasized by folk psychological accounts, or the unreflective, direct perception of people's mental states described above. Heidegger (1927/1962) is an important historical source for this view, but the phenomenological critics also find similar arguments in the work of Aron Gurwitsch (1931/1979),

Maurice Merleau-Ponty (1945/2002), and even Husserl (see Zahavi, 2005, ch. 6). My focus will not be on these historical sources, but on the views put forward by the contemporary phenomenologists when criticizing the folk psychological account.

This second strand of the phenomenological critique of folk psychology is often introduced in the context of criticizing "theory of mind" experiments like the standard false-belief task (Gallagher, 2001, 2005; Ratcliffe, 2007, ch. 4). Folk psychological accounts take the child's situation in such tasks as paradigmatic: they require the child to *observe* someone's behavior, and *predict* her behavior by *attributing* propositional attitudes to her. In a Sally–Anne false-belief task, if shown Sally unsuccessfully looking for the marble in the basket, the child would be expected to *explain* this behavior by saying that Sally wanted the marble and falsely believed that it was in the basket. The phenomenological critics point out that "theory of mind" tasks place the child in the role of *theorist*, providing explanations and predictions of a third-party's behavior. Even if the child is not required to provide verbal explanations and predictions, the child is at least required to somehow report to the experimenter a behavioral prediction, perhaps by pointing to a location. The two dominant folk psychological accounts, TT and ST, attempt to explain precisely this explicit explanation and prediction of behavior based on mental state attribution.

After thus characterizing the nature of the child's cognitive stance in "theory of mind" tasks, the phenomenological critics then propose the following criticism: what is not being tested, nor being explained by TT and ST, about the child's behavior in these tasks is the child's interaction with the experimenter. More generally, the phenomenological critics see the folk psychological picture as failing to adequately

characterize our *everyday, unreflective social understanding*, where we do not explicitly explain and predict people's behavior in terms of their mental states. They see folk psychology as capturing cases where we take a *theoretical*, "third-person" stance on others' behavior, but not necessarily cases where we are a *participant* in social interaction or otherwise unreflectively respond to other people. This descriptive inadequacy of the folk psychological picture, according to the phenomenological critics, also affects the lessons we can draw from empirical research, such as those involving "theory of mind" tasks: since these tasks focus on the reflective or theoretical capacities of explaining and predicting behavior, and do not investigate the socio-cognitive abilities necessary for unreflective social understanding, the theorizing and simulation posited by TT and ST might very well only be restricted to the relatively rare cases of reflective social understanding.

The distinction between theoretical and participatory social understanding to which the phenomenological critics want to call attention can be helpfully understood in terms of Wheeler's (2005) distinction between "online" and "offline" intelligence (which he introduces while developing a Heideggerian conceptual framework for cognitive science). Online intelligence involves an organism's active sensorimotor engagement with the world: "A creature displays online intelligence just when it produces a suite of fluid and flexible real-time adaptive responses to incoming sensory stimuli" (p. 12). Offline intelligence, in contrast, is exhibited when an organism is not acting, but reflecting on the world and its possible actions. This is not primarily a contrast between psychological processes that are explicit and available to consciousness, and ones that are not. Rather, it is about the stance an organism takes

toward its environment: online sensorimotor interaction versus disengaged contemplation.

While standard "theory of mind" tasks require the child to provide a prediction or explanation to the experimenter, and thus involves interaction, traditional folk psychological accounts clearly focus more on *offline* forms of social understanding, where we are *thinking* about other people's behavior and making explicit *judgments* about their mental states—in the case of standard false-belief tasks, using mental state concepts to think about someone's false belief to explain or predict their behavior. What the phenomenological critics want to call our attention to are *online* forms of social understanding, such as the child's active engagement with the experimenter.

The central question raised here by the phenomenological critics is thus: *Does the folk psychological picture capture cases of online social understanding?* Their answer is that it does not. One source of evidence for this claim is phenomenological: we do not often consciously experience making mental state attributions in such cases. If, however, we attend to the distinction between personal and subpersonal levels—i.e., if we distinguish accounts of the phenomenology of social understanding and from accounts of the psychological or neural mechanisms enabling such phenomena—we can call into question the phenomenologists' answer. Their phenomenological evidence certainly cuts against personal-level accounts of our conscious experience during cases of online social understanding. But it does not rule out a role for folk psychology in the neural processes that enable online social understanding. Other sorts of evidence will be necessary to evaluate this possibility. And I will argue that this evidence does suggest a role for mental state attribution in online social understanding.

Turning the focus toward online social understanding does, however, raise a host of new empirical and theoretical issues. Thus, confronting the challenge of the phenomenological critics will help move us toward an enriched account of human social understanding.

3. Synthesizing Folk Psychological and Phenomenological Accounts

I will be arguing in this dissertation that the phenomenological critics have gone too far in their criticisms of the folk psychological picture of social understanding. Instead of replacing the folk psychological account, I will argue that the traditional folk psychological accounts and the insights of the phenomenologists criticizing these accounts can be synthesized into an enriched account of the human social understanding.

Making this argument depends on distinguishing different levels of explanation or analysis at which the phenomenologists' claims are being pitched. The phenomenologists' claims about our experience of other people, and the distinction between online and offline social understanding, concern the *personal level*: i.e., they concern the experiences of and activities performed by persons or agents. It is a very different inquiry to investigate the processes going on inside of agents, particularly their brains, which enable these different personal-level capacities. While TT and ST are often pitched at the personal level, there are versions of each which are subpersonal-level accounts—i.e., accounts of the subpersonal mechanisms which bring about personal-level capacities. Even if the person is not explicitly attributing

mental states, their online social behavior may be produced by subpersonal-level processes which involve representations of people's mental states.

Thus, to properly evaluate the attack on folk psychology, we must have a story about the different levels of explanation for cognitive phenomena, to distinguish competing accounts at the same level from potentially compatible accounts at different levels. This will be the topic of chapter 2. There I will first describe Daniel Dennett's original formational of the distinction between personal and subpersonal levels. Then I will explicate my own characterization of the personal-subpersonal distinction. My account appeals to recent work in the philosophy of science on the nature of mechanisms and mechanistic explanation. I go a bit beyond these discussions by trying to locate a place for phenomenology or conscious experience in multilevel, mechanistic accounts of cognition. With this account of personal vs. subpersonal levels of explanation in hand, I address in chapter 3 the methods of inquiry used to obtain these multilevel accounts of cognitive phenomena. I first describe third-person techniques commonly used in the cognitive sciences, such as behavioral experiments and methods for directly investigating the structure and activity of our neural mechanisms. Since the phenomenologists' critique of folk psychology often appeals to claims about our conscious experience, I also discuss first-person methods—my main example being the method developed in the phenomenological tradition—used to acquire more direct evidence about the structure and contents of consciousness.

In the rest of the dissertation, I apply this explanatory and methodological framework to the debate between the folk psychological and phenomenological accounts of human social understanding. As indicated above, I first address the topic

of direct social perception. In chapter 4, I take on Shaun Gallagher and Dan Zahavi's view that TT and ST are incompatible with these phenomena. I will accept their phenomenological claim about the experiential aspect of direct social perception, while defending the idea that theorizing and simulation remain possible subpersonal explanations of direct social perception.

In chapter 5, I move on to the second major aspect of their critique of folk psychology: their claim that online social interaction does not require folk psychological reasoning involving the attribution of beliefs and desires. There I expose the weakness of the phenomenologists' arguments on this point, and defend the role of belief–desire attribution in unreflective, online social interaction. I focus on what is widely considered the paradigmatic case of folk psychological reasoning: the understanding of other people's false beliefs about the world. I will describe evidence from experimental research that we are able to use false-belief attributions in the context of online social interactions. I will then defend my interpretation of this evidence against possible objections from the phenomenological critics.

In summary, I take the key insights of the phenomenological critics to concern the personal-level phenomena of human social understanding. They identify phenomena which have been underemphasized, left out, or misdescribed by those in the folk psychology camp. But unlike the phenomenologists, I do not believe these phenomena are incompatible with the folk psychological picture. I contend that TT and ST remain viable accounts of the subpersonal-level processes enabling social perception, where we have immediate experience of others' emotions and intentions, and online social interaction, where we are sensitive to others' beliefs and desires

without making conscious judgments about these mental states. I believe this synthesis of the folk psychological and phenomenological accounts enriches our understanding of human social understanding in terms of our descriptions of the relevant phenomena and our explanations of these phenomena.

Chapters 1 and 5 contain material reproduced from "False-belief understanding and the phenomenological critics of folk psychology," *Journal of Consciousness Studies*, *15*(12), 33–56. Permission to reproduce this material has been granted by the copyright owner, Imprint Academic. The dissertation author was the primary investigator and author of this paper.

Chapter 2. Characterizing Personal and Subpersonal Levels

1. Introduction

In the last chapter I introduced two challenges phenomenologists have raised about the role of folk psychology in human social understanding. The first concerns the nature of social perception, and whether folk psychological theorizing and simulation play a role in the recognition of others' emotions and intentions. The second concerns whether mental state attribution is at all required for online social understanding. I suggested, however, that TT and ST can be seen as pitched at a different level of explanation than the level at which the phenomenologists are working, and thus be compatible with their claims about the phenomenology of human social understanding. To make this argument requires filling out what kinds of levels there are in the cognitive sciences, how the levels relate, and the methods and evidence relevant to these different levels. This is an issue that helps not just with my argument about the phenomenological accounts, but is essential for accounts of social understanding generally. For example, simulation is often pitched as a conscious activity persons engage in, but then differentiated from theorizing in terms of differences in cognitive or neural mechanisms. Since the notion of simulation was developed at the personal level, it is important to question whether subpersonal processes really merit description in terms of simulation. The same goes for theorizing. Therefore, we can benefit from clarity about the level at which a particular

account is being pitched, and what one's commitments are about the relations between levels.

In this chapter I will thus provide an account of different levels of phenomena and explanation in the cognitive sciences. The distinction I will be appealing to is the distinction between *personal and subpersonal levels*. These terms are used in a variety of ways, so I must specify the way I will be using these terms here. I will begin with Dennett's (1969) introduction of the distinction, and then show how a few other authors have conceived of the distinction and the relation between levels. My sense of the distinction will depend largely on the notion of *levels of mechanisms* found in the recent literature on mechanistic explanation (e.g., Bechtel, 2008b; Craver, 2007). Descending from the personal level is about decomposing the person into parts, mainly brain parts, and explaining how the organized operations of these parts bring about personal-level phenomena. I will not, however, be taking a stand on precisely how to characterize the relation between conscious experience and bodily (presumably neural) mechanisms. My framework will thus be able to accommodate however this aspect of the mind–body problem ends up being solved.

2. Dennett on Personal and Subpersonal Levels of Explanation

2.1. Dennett's Initial Characterization of the Distinction

When Daniel Dennett introduced the distinction between *personal* and *subpersonal* levels of explanation in his book *Content and Consciousness* (1969), it was in the context of addressing the mind–body problem. His approach was to focus

on our language, on the mode of discourse we use to describe the mind, and its relation to scientific discourse about, e.g., neural and other physical entities. Rather than trying to relate the mental entities described in our mentalistic discourse with the physical entities posited by the sciences—e.g., as with the proposals that mental phenomena are identical to states of the brain, or fail to exist because of a lack of mental–neural identity—Dennett suggested that we suspend any ontological commitments about mentalistic discourse. If we treat mental language as non-referring, we can then ask about what relations hold between the truth of sentences about mental phenomena and the truth of sentences from the sciences, without saying that mental phenomena are identical or non-identical to scientifically reputable entities such as neural states or processes. This methodological context is essential to Dennett's initial marking of the personal and subpersonal levels of explanation.

According to Dennett (1969), our mentalistic and scientific ways of talking use very different vocabularies with unique semantic properties. The vocabulary we use to characterize persons is that of folk psychology (although it was not called that at the time), with terms for psychological phenomena such as beliefs, desires, thoughts, intentions, and actions. As Brentano famously noted, the characteristic feature of mental phenomena is that they exhibit *intentionality* (with a "t") or "aboutness." The way Dennett interprets this is that mental phenomena have content or meaning. Beliefs, for example, are a type of mental state or attitude directed toward propositional contents, which can be expressed linguistically with "that" clauses (e.g., "that the cat is on the mat"). Talk about people's beliefs, desires, etc. thus attributes contentful mental states to persons (e.g., the propositional attitude "Bob believes that

the cat is on the mat"). Such sentences are noteworthy in being *intensional* (with an "s"), i.e., their meaning is not simply determined by the things and properties in the world to which their constituent terms refer. For example, propositional attitude ascriptions do not preserve truth given the substitution of coreferential terms. To use a well-worn case, "Lois Lane believes that Superman can fly" is true, but "Lois Lane believes Clark Kent can fly" is not true, even though "Superman" and "Clark Kent" refer to the same person. Belief and other mental state ascriptions are relative to how people describe or represent entities. The language of science does not include these mental idioms, but instead uses *extensional* terms referring to physical processes, particularly those of the brain and body. These differences between mental and scientific discourse call into question whether they can be reconciled, whether we can find a place in the world for mental phenomena. As noted above, Dennett takes an instrumentalist approach to mentalistic discourse, suspending ontological commitments about it and treating the mentalistic and scientific characterizations as autonomous ways of describing human behavior.

Dennett introduces the notions of *personal and subpersonal levels of explanation* to identify explanations of human behavior using, respectively, these intensional and extensional vocabularies: The personal level of explanation uses mentalistic language, whereas the subpersonal level uses the language of physical events. Related to these differences in vocabulary, Dennett identifies differences in the kinds of explanations provided by these two ways of speaking. While subpersonal-level explanations identify "mechanical" causes amongst the physical entities in the brain and body, personal-level explanations are "non-mechanical" interpretations of

the intentional actions and mental states and processes of rational agents. As Dennett emphasizes in later writings (e.g., Dennett, 1987), during personal-level interpretation of someone's behavior, we posit the mental states (e.g., beliefs and desires) and practical reasoning performed on those states which a rational agent would have been likely to engage in prior to performing the observed action. Such interpretation is thus *holistic*, in requiring the attribution of a set of mental states all at once, and *normative*, in assuming norms such as rationality when attributing mental states and reasoning processes to agents. Personal-level explanation, according to Dennett, is thus a very different explanatory practice than identifying causal processes in the brain and body.

Accordingly, Dennett argues that the kinds of questions asked and the kinds of answers offered at these different levels of explanation must not be confused. When explaining why a person withdraws his hand from a hot stove, the personal-level explanation is simply "that the person has a 'sensation' which he identifies as pain, and which he is somehow able to 'locate' in his fingertips, and this 'prompts' him to remove his hand" (Dennett, 1969, p. 91). We cannot say anything further at the personal level about how a painful sensation is distinguished from one that is not painful, how a pain is located in the body, or what it is about painfulness which prompts one to eliminate or avoid the pain, in this case by withdrawing one's hand. From the personal level, there is nothing more one can say about these activities of persons. There are certain activities and features of persons that are simply primitive, and no further questions and answers can be provided about them in terms of personal-level phenomena. For instance, persons do not distinguish pains from other sensations by noting that they meet certain criteria—they simply experience painful sensations

and can discriminate painful sensations from other sensations. Any further explanation of these personal-level capacities requires that we "abandon the explanatory level of people and their sensations and activities and turn to the *sub-personal* level of brains and events in the nervous system" (Dennett, 1969, p. 93). We can characterize the physical processes in the brain and body during such episodes of pain. But in such explanations we should not, according to Dennett, identify brain processes with personal-level experiences like sensations of pain—we instead "abandon" the vocabulary of mental phenomena and use only the vocabulary of physical phenomena.

The Dennett of *Content and Consciousness* thus sees personal- and subpersonal-level explanations as autonomous, using distinct vocabularies and answering distinct kinds of questions. Personal-level explanations attribute mental contents to persons, while subpersonal-level explanations describe noncontentful physical events. Explanations in terms of the activities and experiences of persons can come to an end, and when we want further answers we must change vocabularies and ask about the causal processes in our brains and bodies. But these different levels of explanation are autonomous. Minimally, this means the two can be characterized independently, such that an account at one level can be offered without requiring an account at the other level. Although some questions which do not have answers at one level may be answerable by switching levels, the precise relation between the phenomena identified at different levels, and the different sorts of questions and answers provided at each level is not well spelled out. For instance, if a personal-level explanation of pain behavior runs out by positing primitive responses of persons, switching levels can allow one to provide a mechanical explanation of a behavior; but

Dennett does not seem to treat this subpersonal account as explaining or accounting for the personal-level phenomena. The subpersonal-level account addresses different questions and phenomena than the personal-level account, and is an *alternative* to a personal-level account rather than accounting for the personal-level phenomena.

Dennett's distinction between personal and subpersonal levels is often characterized in terms of metaphysical levels of composition (e.g., Bermúdez, 2005): that the personal level concerns the activities of *whole* persons, while the subpersonal level concerns how the *parts* of persons are organized. In this way, subpersonal-level phenomena can be seen as constituting or bringing about personal-level phenomena, and thus subpersonal-level accounts as providing constitutive explanations of personal-level phenomena. While Dennett's talk of *levels* of explanation and calling what brains do *sub*personal suggests such a reading, *Content and Consciousness* does not permit such a straightforward reading of Dennett's personal and subpersonal levels. This is because Dennett's distinction does not clearly differentiate two senses of "levels": (1) levels of *nature*, such as levels of composition; and (2) levels of *description* or *analysis*, different descriptions or properties of the same thing (for an extended discussion of different senses of "levels," see Craver, 2007). The distinction between whole persons and parts of persons marks different levels of nature, whose interrelation can be understood in terms of a compositional relation between persons and their parts. But the distinction between intentional and nonintentional explanations is, arguably, a distinction between different ways of characterizing some single thing: using the vocabulary of folk psychology, versus some scientific vocabulary which does not include intentional concepts. Consider two ways a sentence can be analyzed:

in terms of its grammatical structure, and in terms of its meaning. The same thing, the sentence, can be characterized in terms of its grammatical properties (formal features of the sentence) and in terms of its semantic properties (what it is about). Similarly, the person may be analyzed as a physical system using some nonintentional scientific vocabulary, or in the intentional terms of folk psychology. Within cognitive science, David Marr's (1982) famous distinction between computational, algorithmic, and implementational levels of analysis is best understood as providing three ways of describing the same thing, rather than distinguishing levels of nature or phenomena to be described (Bechtel, 1994, p. 9, 2008b, pp. 25–26; Craver, 2007, p. 218). If such descriptions are interpreted as identifying real properties of things, such levels of analysis can be related in terms of *realization* relations: e.g., a cognitive system may exhibit certain algorithmic properties in virtue of having certain physical or implementation-level properties. Read as levels of nature, Dennett's distinction between personal and subpersonal levels would mark the distinction between whole persons and parts of persons. But read as levels of analysis, the same thing, the person, would be characterized in intentional versus nonintentional terms. With this distinction between kinds of levels in place, both whole persons and their parts could in principle be given either kind of description. Dennett, however, does not seem to allow for this possibility in this early work. He seems to restrict intentional vocabulary to the level of whole organisms, and nonintentional vocabulary to the level of neural and bodily parts of organisms. Levels of composition seem to be relevant to Dennett's distinction between personal and subpersonal levels, but are not adequately distinguished from levels of realization to be all that Dennett intends.

Even this minimal reading of the role of levels of composition in Dennett's personal–subpersonal distinction might be inadequate, however, since it's not clear that Dennett (at least in 1969) thinks we can truly talk about "person parts." When Dennett writes of "abandoning" personal-level phenomena such as pains when one moves to the subpersonal level, it's not clear that this is analogous to moving from talk about a whole and its properties to talking about parts of the whole, their properties and how they are organized (e.g., no longer talking about a whole car and its properties when you start talking about the car's parts and their properties). First, and most importantly, Dennett is basically instrumentalist about our talk at the personal level, treating intentional descriptions as non-referring. So persons and their brains cannot be linked by a relation of composition, since the personal-level phenomena we talk about are not given ontological import. If this were not enough, explanations at the two levels differ so significantly that it is difficult to treat their relation as one between explanations at different levels of composition. As Hornsby (2000) argues, the "non-mechanical" explanations in terms of content offered at the personal level are holistic and normative. This is what makes them "personal," in contrast to the "impersonal," non-normative identification of non-contentful, physical causes found in subpersonal "mechanical" explanations. Dennett does not explicitly describe this unique character of personal-level phenomena as being constitutively explained by subpersonal-level phenomena, like the properties of a whole can be constitutively explained by the properties and organization of its parts (see Hornsby, 2000). So even if Dennett were more realist in his treatment of the personal level, the very different explanations found at the personal and subpersonal levels would suggest that (a) the

phenomena posited at the two levels cannot be easily characterized in terms of a compositional relationship, and (b) subpersonal-level explanations cannot obviously be treated as reductive, constitutive explanations of personal-level phenomena.

Dennett's personal and subpersonal levels are thus more strongly autonomous than would be suggested by treating the distinction as simply marking explanations appropriate to wholes versus those appropriate to parts of wholes. Although distinguishing the activities of persons from the activities of brains is an important part of the personal–subpersonal distinction, precisely how the two levels are related metaphysically and explanatorily is not especially clear in *Content and Consciousness*. This is likely due to Dennett's explicitly instrumentalist stance on personal-level phenomena.

2.2. Developments in Dennett's Treatment of Personal and Subpersonal Levels

Of course Dennett has written a tremendous amount since 1969, developing and arguably changing his views on the nature of intentional and other forms of explanation, and how these phenomena and explanations interrelate. I can only touch on a few themes that are relevant to how others have used Dennett's personal–subpersonal distinction, and how I will be marking it.

First, Dennett becomes more comfortable in later writings with treating subpersonal-level accounts as providing reductive, constitutive explanations of personal-level phenomena: "Sub-personal theories proceed by analyzing a person into an organization of subsystems…and attempting to explain the behavior of the whole person as the outcome of the interaction of these subsystems" (Dennett, 1981, p. 153;

as cited in Hornsby, 2000, p. 16). While maintaining that questions and answers at different levels must not be confused, Dennett's later writings more clearly treat persons as wholes whose activity can be constitutively explained by decomposing the system and analyzing its parts and their organization—rather than understanding personal- and subpersonal-level accounts as autonomous, alternative descriptions, where the relation between phenomena across levels is left open.

Further, as Hornsby (2000) notes, Dennett begins allowing attributions of content to subpersonal brain mechanisms, so that content is no longer exclusive to the level of whole persons. This is explicit in Dennett's explanatory strategy of "homuncular functionalism," expressed in the above quotation. The person is first decomposed into a set of intelligent subsystems, each given an intentional interpretation of how they accomplish their input–output function. The subsystems are then each decomposed into "dumber" subsystems, with the decomposition continuing until it bottoms out in activities simple enough to be performed by unintelligent physical mechanisms. The strategy does not suffer from an infinite regress of homunculi because the treatment of subsystems as intentional systems is eventually cashed out in terms of nonintentional processes. This treatment of subsystems as intentional systems is made possible by Dennett's explicitly pragmatic approach to attributions of content: intentional explanation is useful for characterizing the input–output behavior of some systems, and its usefulness is determined by its predictive and explanatory success, not the kind of thing to which it is being applied. So if the intentional stance is helpful in understanding both persons and parts of persons, so be it. This pragmatism about intentional ascriptions can be seen as a natural extension of

Dennett's suspension of ontological commitments for personal-level explanations in *Content and Consciousness*. If ascribing intentional states is a matter of interpretation, there is not a principled reason for restricting these interpretations to any particular kind of physical entity, as long as such interpretations are useful.

With this allowance of subpersonal ascriptions of content, Dennett also expands the range of intentional concepts used beyond those of folk psychology. In "Three Kinds of Intentional Psychology," Dennett (1987, ch. 3) characterizes "subpersonal cognitive psychology" as a performance model of the *brain's* operations characterized in terms of the manipulation of content-laden intentional states. While maintaining that subpersonal cognitive psychology will appeal to intentional concepts, Dennett claims that it might require new intentional concepts which significantly depart from our commonsense or folk psychological concepts and only resemble the folk concepts in being intentional (p. 63). So not only does Dennett accept ascriptions of content to subpersonal brain parts, but he thinks this will involve conceptual progress with regard to the types of intentional states being ascribed. Dennett thus gives no privilege to the intentional vocabulary found in folk psychology (at least the mentalistic discourse found in Western culture in the twentieth and twenty-first centuries). This is significant because some theorists reject the notion of subpersonal content entirely, maintaining Dennett's original treatment of intentional vocabulary as purely at the level of whole persons/organisms, while others seek to sharply distinguish personal- from subpersonal-level content.

With this move toward characterizing the inner parts of organisms as content-bearing states, Dennett does, however, maintain personal-level intentional attribution

as an explanatory practice independent of commitments about inner states. Dennett's ontological commitments about mental states are still tricky to pin down, leaving him open to charges of instrumentalism about mental states. In response to such charges, Dennett repeatedly affirms that he is a realist of some sort about beliefs and other mental states. Yet what he treats intentional attributions as picking out are not features of people's internal workings (Dennett, 1987). Instead, mental state attributions pick out "real patterns" in the behavior of agents (Dennett, 1991b). Clark (1993) exposes a tension in Dennett's account between treating intentional stance ascriptions as completely uncommitted about the inner workings of persons, and making some commitments about how persons are organized internally. Clark (1993) and Bechtel and Abrahamsen (1993) offer a realist middle ground, where folk psychological attributions mainly characterize an agent's relation to her environment, what she is informationally sensitive to and how she is disposed to behave. This puts some constraints on what a person's inner processes must be like, in order to enable informational sensitivities and actions of certain types. But folk psychological attributions need not make specific claims about the internal mechanisms by which these general information-processing activities are accomplished. While Dennett does not explicitly adopt this approach, it is one he recognizes as open to him—and for many, is more attractive than Dennett's own talk of real patterns.

In summary, Dennett's more recent writings better distinguish levels of nature from levels of analysis. With regard to levels of nature, he seems to treat whole persons as composed of parts. And by allowing that neural processes can be described both nonintentionally and in terms of their content, Dennett makes a distinction

between different levels of analysis of the same thing. It is hard to tell, however, exactly how these two notions of levels relate to his distinction between personal and subpersonal levels. Since my focus in this dissertation is not interpretation of Dennett's writings, I will move on from Dennett and sketch my account of personal and subpersonal levels.

2.3. Personal and Subpersonal Levels as Levels of Mechanisms

From this survey of Dennett's work, we can see that the distinction and relation between personal and subpersonal levels can be conceived of in many ways. But what I see at the heart of this distinction is the idea that minded beings are complex systems with a hierarchical organization, and that different forms of explanation are appropriate at these different levels of nature. The activity of brain parts must be distinguished from the activity of whole persons, and the goal of the cognitive sciences is to determine how personal-level phenomena such as perceiving, remembering, thinking, imagining, etc., are produced by the organized activity of the neural and other physical processes constituting the human agent.

From this core distinction, we are left with many open questions about the nature of these levels and how they relate. Whole books are written summarizing the various positions in the literature on the appropriate metaphysical and explanatory levels in the cognitive sciences (e.g., Bermúdez, 2005). And of course I cannot provide a detailed picture of how these questions should be answered. These are controversial issues, and I do not want my work on the nature of social understanding to be firmly committed to a particular view on these issues—though given the range of

views in the literature, I will surely disagree with some positions in order to say much at all with regard to these questions.

My approach will be to read the personal–subpersonal distinction through the recent work on mechanistic explanation (e.g., Bechtel, 2008b; Craver, 2007). Mechanisms are conceived of as complex entities composed of parts. At the level of the whole mechanism, *etiological* explanations can characterize how environmental phenomena affect the activity of the mechanism. Etiological explanations can also be offered at a lower level to characterize the causal interactions among the mechanism's parts. These metaphysical levels are related in that the activity of a mechanism is *constituted* by the organized activity of its parts, rather than being caused by the activity of the parts. Mechanistic explanations are reductive in that they investigate the operations of the parts of mechanisms. But in doing so, they do not eliminate higher-level phenomena in favor of lower-level phenomena. How working parts are organized to constitute the whole mechanism is a crucial part of mechanistic explanation. Further, mechanistic explanations can involve a hierarchy of levels, in that the parts of a mechanism can be mechanisms and decomposed into their parts. Mechanistic levels are not characterized in terms of the size of components, or the kind of entities specific to particular sciences, but are relative to the decompositional analysis of a particular mechanism performing a particular function.

Read in light of the mechanistic approach, whole persons or agents are conceived of as mechanisms that engage with their environments. The personal level will involve characterizing the activities and experiences of whole persons in their physical and social environments. The person must however be decomposed to

determine how the organized activity of its parts bring about these personal-level phenomena of perceiving, remembering, thinking, imagining, etc., when the whole person is properly engaged with their environment. Rather than a single subpersonal level, there will be a hierarchy of subpersonal levels as the person is decomposed into parts, and these parts themselves decomposed, and so on. The mechanistic approach does not take an a priori stand on how subpersonal processes are to be characterized—e.g., whether it be in terms of a Fodorian language of thought, connectionist networks, or some as yet to be discovered vocabulary for cognitive operations. The key emphasis is on determining how the actual biological mechanisms of cognition operate. While theorists may come up with various models of how a system may possibly be organized, mechanists want to understand how real organisms are constituted, and this will require the tools of neuroscience to determine how the brain is organized and what its parts do. The behavioral experiments of psychology thus must be integrated with other forms of inquiry into how neural activities function to bring about personal-level behavior.

With the mechanistic framework, we have a coherent way of talking about metaphysical and explanatory levels relevant to the cognitive sciences, and one which is based in the way practicing scientists attempt to explain cognitive phenomena, from the behavior of whole agents all the way down (at least) to the molecular composition of neurons. Of course there remain many open questions about how to provide integrated, interlevel accounts of cognition within this framework. But I think the mechanistic framework provides a useful way of understanding the distinction

between personal and subpersonal levels, without the baggage of some other characterizations of it.

In line with Dennett's more recent writings, the mechanistic approach can allow contentful and noncontentful attributions at both personal and subpersonal levels. The mechanistic approach need not take an a priori stand on how content and the personal and subpersonal levels are related. Although it is common to identify personal-level content with the content of neural states, this assumption is not uncontroversial (see, e.g., Noë & Thompson, 2004). While the mechanistic approach sees such identifications as methodologically useful heuristics for providing constitutive explanations of higher-level phenomena in terms of lower-level phenomena, it need not treat personal-level content as simply identified with subpersonal-level content.

With this general approach now sketched, I will address a few more important points about the personal and subpersonal levels made by John McDowell and Susan Hurley. While I do not agree entirely with either author's treatment of levels, each addresses issues central to my characterization of the personal and subpersonal levels.

3. McDowell on the Personal–Subpersonal Distinction: Phenomenology and Interlevel Relations

In "The Content of Perceptual Experience," John McDowell (1994) critiques an early paper by Dennett, "Toward a Cognitive Theory of Consciousness" (1981, ch. 9), with regard to its personal- and subpersonal-level accounts of perceptual

experience. Without going into detail about either Dennett's account or McDowell's critique, I want to call attention to a few points McDowell makes with regard to the personal and subpersonal levels: (1) using phenomenological adequacy as a criterion for personal-level accounts of mental content; and (2) his distinction between "constitutive" and "enabling" explanations associated with, respectively, the personal and subpersonal levels.

3.1. McDowell on Phenomenological Adequacy

One of the first issues McDowell addresses is the *phenomenological adequacy* of Dennett's account of perceptual experience. According to McDowell, accurate phenomenological description is a key criterion of personal-level accounts, and one must not infect personal-level accounts with features of subpersonal-level accounts. McDowell identifies both accurate and inaccurate features of Dennett's treatment of the phenomenology of perceptual experience.

McDowell starts by praising Dennett's remark that we have "no direct personal access" to the structure of our subpersonal vehicles of content (p. 190). The remark turns on a distinction between properties of *content* and properties of the *vehicles* carrying that content. Dennett here mentions the debate about whether perception involves propositional or imagistic representations. He notes that this is a debate about the vehicles of content, which has nothing to do with the particular content carried by these vehicles. Dennett then asserts that what is experienced at the personal level is only the content and not the vehicle or structural properties of these subpersonal representations. McDowell treats this as an accurate description of our personal-level

phenomenology. For example, he asserts that our visual experience is not of images, but of the relevant parts of the environment, even if subpersonal vehicles are indeed image-like in structure. So, according to McDowell, Dennett correctly avoids infecting his personal-level account of the content of conscious experience with features of his subpersonal-level account of representational vehicles.

Although he agrees with Dennett on this point, McDowell diagnoses a personal–subpersonal or content–vehicle confusion in Dennett's treatment of perceptual experiences as "presentiments." The notion of "presentiment" captures experiences where you simply have the experience without knowing why (i.e., without knowing its causal origin), such as being struck by the thought that someone is looking over one's shoulder as one writes. Dennett believes perceptual experiences are also like this: that a perceptual experience is a hypothesis generated by various other subpersonal processes to which we do not have access, and since the content of our experience contains nothing about these causal processes responsible for generating it, we experience it as a presentiment. McDowell objects to Dennett's treatment of visual experience in terms of presentiments, arguing that the phenomenology of visual experience does indeed include an experience of the causal origins of that content—that a perceived object is *experienced as* present in one's awareness, rather than the presence of the object being experienced as a hypothesis. McDowell diagnoses this inaccurate description as being due to a personal–subpersonal or vehicle–content confusion: that Dennett's account of the subpersonal processes underlying conscious experience leads to an inaccurate account of the personal-level phenomenology.

Whether McDowell is correct or not with regard to these arguments, McDowell identifies what I take to be a key criterion for accounts of conscious personal-level content: namely, that they accurately describe the phenomenology of experience, or how things appear to us. Although not all personal-level mental states are conscious, and it is sometimes unclear whether contents labeled "unconscious" are posited at personal or subpersonal levels, conscious experiences are paradigmatic examples of personal-level mental states. Accordingly, the phenomenology of conscious experience must be accurately described. Note, however, that McDowell's phenomenological claims are not about qualia, but about the *content* of experience. Although the qualitative features of conscious experience are an essential component of any account of conscious experience, phenomenological claims are not exhausted by claims about qualia. Some will argue that phenomenal character just is a matter of representational content. But without taking a stand on the nature of the qualitative character of experience, it is important to note that phenomenological description is as much about the content of experience as its qualitative character. In fact, the philosophical tradition of phenomenology has been particularly concerned with the intentional content of experience, even if in analytic philosophy the term "phenomenology" has usually been restricted to the qualitative features of experience.

How we actually go about acquiring the kind of phenomenological descriptions McDowell uses in his critique of Dennett is of course a very difficult issue. I will say more about first-person methods for describing conscious experience in chapter 3. But I generally will be remaining uncommitted about these issues. I want to note, however, a potential conflict between the phenomenological characterization

of personal-level content, and the "interpretative" approach to personal-level content found in the writings of Dennett, Davidson, and others. Dennett (1987) treats personal-level attributions of mental content to be a matter of interpreting behavior, even going so far as to say that mental state attributions can be indeterminate—e.g., that multiple sets of beliefs and desires may equally predict and/or explain a person's behavior, such that we have no means of picking between them. This third-person approach to mental content attribution seems, at least on its face, in tension with the first-person phenomenological approach. For example, from the first-person perspective the intentional content of one's experience does not seem capable of the radical indeterminacy Dennett describes. This issue of indeterminacy is just one reason people are skeptical of the interpretative approach, and hope to provide a more realist, determinate account of mental content. One proposed solution is to look to determinate internal states of organisms and try to identify the contents of these. Of course the interpretative approach does not describe only conscious mental states, for a behavior can be explained in terms of mental states of which the person has no first-person awareness. And a first-person approach will provide no evidence of *unconscious* mental content (if a notion of personal-level unconscious mental states makes sense).

These are very tricky issues, and I do not have the space to address them adequately. I will just note what I think is most essential for my purposes. By raising the issue of phenomenological adequacy, McDowell identifies a very important criterion for personal-level accounts. Of course phenomenological considerations have no purchase with regard to unconscious personal-level states (assuming this remains a

useful concept once the personal–subpersonal distinction is in place). But with regard to some accounts, identifying what we are *not* conscious of is an important corrective. And one must tell a story of how to go about attributing unconscious mental states, and how to conceive of the relation between conscious and unconscious personal-level states. Yet consciousness is a paradigmatic feature of the personal level, and first-person approaches to characterizing experience must not be forgotten in the scientific study of mind. Further, phenomenological accounts must be recognized as at the personal level, and distinct from claims about brain processes, although of course these are importantly related. Saying something about how they are related will be the topic of the next section.

3.2. McDowell on "Constitutive" vs. "Enabling" Explanations

As the phenomenological considerations should make clear, McDowell draws a firm line between personal- and subpersonal-level accounts. The personal level concerns whole persons "more or less competently inhabiting an environment," which includes our being "informed of" features of our environment through our senses (p. 196). This is what talk of personal-level content characterizes: the engagement of persons with their environments. Such accounts McDowell calls "constitutive explanations" of personal-level phenomena. In contrast, the subpersonal level is, for McDowell, the level of noncontentful, syntactic processes. Even if it is useful to characterize subpersonal processes as carrying content, these are not genuine attributions of content according to McDowell, since subpersonal processes are syntactic rather than semantic. Personal- and subpersonal-level accounts must not be

confused. But while they are autonomous, McDowell does treat the two levels as interrelated, rather than simply two wholly independent alternatives. He sees subpersonal accounts not as explaining what personal-level phenomena *are*, but as providing "causal" or "enabling" explanations of the physical processes "in virtue of which" we have personal-level content. Subpersonal processes do not "constitute" personal-level content, since there is no content at the subpersonal level. But subpersonal accounts "make intelligible" personal-level phenomena; without the proper organization and operation of our neural mechanisms, we would not have the conscious experience we exhibit.

Using this distinction between constitutive and enabling explanations, McDowell criticizes Dennett for his characterization of the relation between subpersonal mechanisms and personal-level content. Dennett's view is that the content of conscious experience can be constitutively explained in terms of a subpersonal mechanism for linguistic output accessing some of the content carried by other subpersonal mechanisms. McDowell allows that Dennett's account of subpersonal mechanisms, removed of its claims of genuine content, may be an accurate characterization of how neural mechanisms causally interact, and that this is what "enables" or provides the "causal" basis for conscious experience. But with only this subpersonal account of consciousness, "We lack an account of what it is, even if we have an account of what enables it to be present" (p. 203). Since there is no content at the subpersonal level, according to McDowell, personal-level content cannot be constitutively explained in terms of the access a subpersonal mechanism has to other subpersonal mechanisms. To talk of conscious experience being a matter of accessing

the content of internal, subpersonal mechanisms is to confuse personal and subpersonal levels. We do not, from the personal level, have any access to our subpersonal-level mechanisms as such. A constitutive explanation of conscious experience, say perceptual experience, requires characterizing our engagements with the world, not our interiors. Inner, subpersonal processes enable this experience, but does not constitute it.

This description might give the impression that McDowell sees personal- and subpersonal-level explanations as completely autonomous from one another. But McDowell is explicit that there can be mutual interaction between accounts at personal and subpersonal levels. He illustrates this with an example from Dennett: the attempt to characterize the perceptual experience of frogs.

> Casual observation of frog life might induce the provisional thought that frogs become informed, through vision, of the presence of bugs. Then it turns out that a good theory of the relevant perceptual equipment fails to support the view that the equipment processes information about arrays of light into information about the presence of bugs. The equipment hardly processes information at all (it is a limiting case of an information-processing device), but rather simply reacts to any small moving speck. It is better to view the informational transaction as the transmission, to "motor control," of information to the effect that a small moving speck is at such and such a point in motor space. So we recast our conception of what frogs become informed of: at best the presence of a bug-like object at a certain place. (p. 196)

McDowell treats the talk of subpersonal information processing, of subpersonal mechanisms "telling" each other things, as metaphorical compared to the genuine notion of persons communicating content to one another. But he accepts that the attribution of "as-if" content to subpersonal mechanisms can inform one's account of personal-level content. Here, studying the visual systems of frogs causes a revision in

the personal-level account of their perceptual experience. But while he accepts there can be interaction between different levels of explanation, McDowell insists that the two levels must be kept distinct:

> The fact that there is this perfectly intelligible interplay between what we decide we can correctly say, in content-involving terms, about frogs, on the one hand, and the detail of a content-involving (information-processing) account of the inner workings of the parts of frogs, on the other, is no reason to mix the two stories together. (p. 197)

How precisely to characterize the influence accounts at different levels have on one another is, however, not very clear. McDowell's distinction between constitutive and enabling explanations suggests the explanatory autonomy of personal- and subpersonal-level accounts. One can offer explanations fully in intentional terms at the personal level, or fully in (nonintentional) neuroscientific terms at the subpersonal level.[9] As accounts of the phenomena at these different levels, they seem explanatorily autonomous. The claim that subpersonal-level phenomena "enable but do not constitute" personal-level phenomena does not tell us very much about the nature of the "enabling" relation. Bermúdez (1995) criticizes McDowell on this issue of interlevel relations, calling into question his use of the semantics–syntax distinction to mark autonomous personal and subpersonal levels, and for in general treating subpersonal-level explanations as only providing enabling conditions of personal-level content attributions.

---

[9] Although I repeatedly contrast the personal level with the subpersonal "level," it must not be forgotten that on my mechanistic view, there will be various subpersonal levels as a mechanism is broken up into parts, and those parts are decomposed, and so on.

Although McDowell's treatment of personal and subpersonal levels in terms of the semantic and syntactic properties of formal languages has its problems, I think there is something right to McDowell's treatment of personal and subpersonal levels as autonomous. Saying how it is right requires, however, a different reading of the metaphysics of these levels. If the personal level is understood as a level of nature consisting of whole organisms, and subpersonal levels as consisting of entities which are parts of organisms, we can understood the distinction as one between levels of composition. This means explanations at different levels are explanations of different phenomena, rather than different explanations of the same phenomena. But these levels of nature can then be explanatorily related in a way pointed to by McDowell's talk of "enabling" relations. The *phenomena* at these different levels may be metaphysically distinct, but to *explain* a personal-level phenomenon might involve descending levels to talk about the organization and activities of the organism's parts. According to this mechanistic conception of levels, the personal and subpersonal levels would thus involve distinct phenomena, but be metaphysically and explanatorily interrelated. Explanations at the level of parts would indeed help to explain phenomena at the level of wholes, but there could be explanations to be had at each level without noting their relations. This is because many mechanists restrict causation to intralevel relations, and talk of "constitutive" relations between levels (Craver & Bechtel, 2007). This contrasts with McDowell's terminology, where "constitutive" explanations are intralevel, and interlevel "enabling" relations are seen as causal. But by treating personal and subpersonal levels as levels of mechanisms, one gets a clearer account of the metaphysical and explanatory relations between levels. The level of

persons and the level of brain parts are metaphysically distinct levels of composition. Etiological, causal explanations can be offered at each level, and constitutive explanations can characterize how the organization of parts can explain/enable the activities of whole persons.

Thus, while I do not endorse McDowell's particular conception of personal and subpersonal levels and their relation, its basic structure is fairly close to the way mechanists offer multilevel explanations. Both maintain the existence of distinct metaphysical levels, with explanations appropriate to each level, while allowing interlevel explanatory relations.

I have suggested that the mechanistic sense of levels works quite well for explaining people's environmentally situated behavior in terms of their inner parts. But does this extend to the other personal-level phenomena McDowell emphasizes, to phenomenology? Does it make sense to treat people's conscious experiences as properties of whole persons the same way behaviors are, which are constitutively explained in terms of brain mechanisms? Here I think the appeal to levels of composition breaks down.[10] It does seem to make sense to ascribe conscious experiences to whole persons—i.e., to treat both behaviors and experiences as at the personal level. But the relation between conscious experiences and brain processes (assuming it is in fact the brain that enables conscious experience[11]) does not seem

[10] Thanks to Carl Craver for helpful discussion on this point.

[11] According to some proponents of embodied cognition, the nonneural body may play a role in enabling conscious experience. I will discuss below some of the issues raised by extended, embedded, embodied views of cognition.

well understood in terms of a part–whole relationship. While there is certainly no consensus in the literature on the metaphysics of mind, some other kind of relation, such as supervenience or realization, is probably more appropriate in this case. So moving "up" from the level of the brain to the level of conscious mental states requires departing from the mechanistic, compositional sense of "levels." McDowell's notion of an "enabling" relation between levels is useful as a neutral term for whatever this relation ends up being. For now, we can say that the brain enables conscious experience, without specifying the nature of this "enabling" relation.

It would be nice if we could invoke a single sense of levels in cognitive science. Unfortunately things are not so simple. The mechanistic sense of levels, where levels are defined in terms of compositional relations, works quite well to explain the behavior of organisms or persons in terms of brain parts, which are further decomposed. Of course each of these levels may be described in various ways—for example, in terms of Marr's three levels of analysis. But the notion of mechanistic levels gives us a unified metaphysical relation for explanations in the cognitive sciences—at least when we put aside consciousness. When the personal-level phenomena are experiences rather than behavior, however, it does not make sense to relate these phenomena to the brain (or possibly the brain plus aspects of the nonneural body) by a compositional relation.

3.3. Lessons from McDowell

What we can take away from this discussion of McDowell's work is the following. Phenomenological description of conscious experience can provide an

important criterion for personal-level accounts, and one which subpersonal accounts must respect in order to offer explanations of these personal-level phenomena. In doing so, we must not confuse phenomena at these different levels, or the notion of levels at issue. For example, we must not mistake attributions of content to brains states for attributions of content to persons, and vice versa. How conscious experience is to be enabled by contentful subpersonal processes remains a matter of controversy. But even without a complete interlevel story here, it is important to keep these levels distinct.

Further, phenomenological accounts can provide important correctives to subpersonal accounts, in providing a better "task analysis" of the personal-level phenomena which subpersonal process are to constitutively explain (Pessoa, Thompson, & Noë, 1998). McDowell's frog example shows a bottom-up influence from the subpersonal level to personal level. But phenomenological descriptions can similarly exert top-down pressure on subpersonal-level accounts. As Alva Noë and Evan Thompson (Noë, 2004; Noë & Thompson, 2004; Pessoa et al., 1998; Thompson, Noë, & Pessoa, 1999) have argued, our perceptual experience is often much less detailed than some researchers assume; this means subpersonal representations need not be as detailed as has been assumed. The brain need not do a lot of "filling in" to provide detail to our visual experience which is not really there; the sense that our visual experience is detailed is, arguably, due to our ability to *access* this detail at will by moving our bodies appropriately. As this case shows, a proper phenomenological description can affect tremendously the assumptions of researchers attempting to offer constitutive subpersonal explanations of personal-level phenomena.

But in constructing interlevel accounts, we must be careful not to assume personal and subpersonal levels must be related by strict isomorphisms (Dennett, 1991a; Gallagher, 1997; Hurley, 1998; cf. Wheeler, 2005). That is, a personal-level phenomenon with a particular phenomenological structure need not be enabled or constituted by subpersonal, neural processes with exactly that structure. For example, the temporal relations of neural processes need not mirror the temporal order of phenomenological experience (Dennett, 1991a; Gallagher, 1997; Grush, 2006; Hurley, 1998). The neural processes might have a much more complicated temporal structure than the serial ordering of personal-level experience. Further, the neural processes enabling a single, unified experience may be highly distributed across the brain—phenomenological unity does not require spatial unity of subpersonal process. What is needed is some intelligible story about how the subpersonal processes enable the personal-level experience. But these cases of more complicated interlevel relations do not rule out isomorphism as one way in which subpersonal-level phenomena can explain personal-level phenomena. We just cannot let an assumption of interlevel isomorphism bias us against other possible interlevel relations, or bias our investigation of the phenomena at each level.

Although I have emphasized the importance of phenomenology to personal-level accounts, we must not overplay the role of conscious experiences in our daily lives. Although we currently lack good stories about how conscious experience arises from subpersonal processes (neural or otherwise), and about the precise role of consciousness in the production of action, we do know that much of our bodily activity is driven by subpersonal processes of which we have no conscious awareness.

We often are not aware of the environmental stimuli affecting our decision-making and behavior, or even that cognitive processes have taken place (see, e.g., Nisbett & Wilson, 1977). While the phenomenological critics emphasize what can be said about conscious experience through first-person methods, they do not deny this point about the need to appeal to subpersonal-level phenomena in constitutive explanations of human behavior, as I will show in section 5. In both chapters 4–5 I address subpersonal-level forms of folk psychological reasoning occurring outside of conscious awareness.

A final point about phenomenology. The phenomenological resources I will be appealing to go beyond those used by McDowell. As seen in chapter 1, I will be following philosophers like Shaun Gallagher, Dan Zahavi, Matthew Ratcliffe and others in appealing to the phenomenological tradition of Husserl, Heidegger, Merleau-Ponty, etc., for descriptions of the personal-level phenomena for which we must provide subpersonal constitutive explanations. This is precisely the strategy of Wheeler (2005) in his development of a Heideggerian conceptual framework for cognitive science, which I appealed to in chapter 1. Indeed, Wheeler explicitly notes the similarities between his Heideggerian multilevel approach and that of McDowell. I will say more about the methodology of obtaining phenomenological descriptions, in the phenomenological tradition and contemporary experimental research, in the next chapter.

4. Susan Hurley on Personal and Subpersonal Levels: Vehicle Externalism

There is one further point about personal and subpersonal levels which I will explicate through the work of Susan Hurley (1998, 2003a, 2003b, 2005, 2006, 2008). Hurley distinguishes personal and subpersonal levels in terms of the vehicle–content distinction: personal-level mental content is "enabled" (in McDowell's sense) by subpersonal vehicles, which can be characterized functionally or in terms of the physical (often neural) processes which implement these functions. Hurley thus endorses a tripartite classification of levels: personal, subpersonal functional, and subpersonal implementation levels.

Hurley (2008) explicitly describes these as "levels of description," which differs significantly from my characterization of personal and subpersonal levels as distinct levels of nature. I am not concerned here, however, with explicating her taxonomy of levels and comparing it to the mechanistic account I have adopted. Rather, my focus will be on a thesis she endorses about the relation between personal-level mental content and the subpersonal vehicles of content: what she calls *vehicle externalism* (see, e.g., Hurley, 1998). This is the idea that subpersonal vehicles need not be restricted to processes inside an organism's brain or even body, but can extend out into the world beyond the organism. Hurley's vehicle externalism is motivated by her more general attack on confusions of personal-level mental content with subpersonal-level vehicles, and assumptions about isomorphisms between levels. For example, Hurley questions the assumption that personal-level conceptual thought must be subserved by language-like subpersonal vehicles inside the brain. For Hurley, it is an open question whether conceptual structure at the personal level requires this same

structure at the level of vehicles, and whether vehicles must be located inside the person rather than extending out into the world.

Beyond Hurley's presentation of it, the general idea of vehicle externalism—which goes by many names, including "active externalism," "the extended mind" (Clark & Chalmers, 1998) "distributed cognition" (Hutchins, 1995), and "enactivism" (Thompson, 2007)—is relevant because it is seemingly in tension with the mechanistic, compositional reading of the relation between personal and subpersonal levels which I have been offering. If the subpersonal substrate of personal-level cognition goes beyond the skin, the personal level cannot be metaphysically related to the subpersonal level by a relation of composition, since this implies the higher level spatially contains the lower level entities—or, it requires drawing the boundaries of the person much wider than is traditionally done.

Although there is much to be said about how one should draw boundaries around cognitive systems, for now I will simply note that the basic idea behind vehicle externalism need not be seen as in tension with the mechanistic approach (Bechtel, 2009; Clark, 2007). That is, interlevel mechanistic explanations give an important role to characterizing how a mechanism, in this case, an embodied agent, engages with its environment. But this does not require that we redraw the boundary of what constitutes the person to include features of the environment. We can still treat the activities of components within the organism as constitutively explaining personal-level phenomena that involve intense organism-environment interaction. I will not here address arguments for this conclusion, such as Bechtel's (2009) discussion of organisms as autonomous systems (cf. Grush's, 2003, argument based on the "plug

criterion" that the *brain* is a self-contained system for the purposes of cognitive science). Note, however, that the mechanistic approach does not rule out the existence of phenomena for which supra-personal levels of analysis are appropriate. For example, some social interactions may be complex enough to characterize the social group as an entity composed of individual organisms.

I have not yet specifically addressed conscious experience in this discussion of vehicle externalism. As I indicated above, I do not think conscious phenomena are related to the brain by the kind of compositional relation holding between behavioral phenomena and the brain. But I adopt the same general perspective on vehicle externalism about both types of personal-level phenomena.

I will put aside the thesis of vehicle externalism, and leave us with the idea that organism–environment relations are essential to personal-level accounts which are then complemented with decompositional, constitutive explanations of how a person's inner parts act to bring about personal-level phenomena. The mechanistic framework treats these two types of analyses as complementary parts of multilevel mechanistic explanations.

5. Summary of Personal and Subpersonal Levels

Let me summarize where we have come thus far. The distinction between personal and subpersonal levels was first explicitly characterized by Dennett. Although it was developed in the context of an instrumentalist approach to mental discourse, a core idea that has come to be associated with these terms is that persons or

organisms are hierarchically structured entities, which can be analyzed at different levels of composition—the personal level of whole persons and various subpersonal levels as the person is decomposed into parts. Whole persons can then be constitutively explained in terms of the operations of their organized parts. Personal and subpersonal levels are thus not simply alternative descriptions of the same phenomena, but aimed at different levels of nature. Adequate accounts of psychological phenomena will thus be interlevel accounts, characterizing both personal-level phenomena and the subpersonal-level processes by which they are enabled. Although I disagreed with his characterization of these levels and their relation, I used McDowell's account of constitutive versus enabling explanations as a foil to characterize the mechanistic view of interlevel accounts.

I also used McDowell to emphasize the use of phenomenological descriptions in personal-level accounts. Phenomenological descriptions can help characterize the personal-level phenomena for which subpersonal processes must account. But first-person approaches are not the only source of evidence for personal-level accounts. As seen in McDowell's example of visual neuroscience influencing accounts of the visual experience of frogs, subpersonal accounts can exert bottom-up influence on personal-level accounts of mental content. More directly, behavioral experiments provide data about personal-level phenomena, such as how persons respond to certain stimuli in particular contexts.

The theme of avoiding interlevel confusions is one present in all three of the authors I've discussed in this chapter. I introduced Hurley's vehicle externalism as an example of just how far apart some try to pull personal and subpersonal levels. In the

context of this dissertation, vehicle externalism plays the role of (a) emphasizing organism–environment interactions, which the mechanistic framework itself emphasizes as an essential higher-level feature of their interlevel accounts; and (b) opening up the possibility of going beyond the organism for explanations of social interaction. For the purposes of this dissertation, this latter possibility will not be emphasized. But the importance of studying organism–environment interactions will be very important to characterizing the phenomena to be explained by subpersonal accounts.

While not wholly uncontroversial, I believe this picture of the personal and subpersonal levels is a common one, and given its basis in mechanistic philosophy of science, is consistent with contemporary research in the cognitive sciences. In the next chapter I will address the research techniques used in the cognitive sciences to obtain evidence about phenomena at the personal and subpersonal levels. Before moving to these methodological issues, however, it is worth noting that even the phenomenological critics appear to share the conception of personal and subpersonal levels I have sketched above. For example, Gallagher and Zahavi (2008) similarly seem to treat personal and subpersonal levels as metaphysically distinct but related levels of nature. They follow McDowell in referring to subpersonal level processes as "the internal enabling conditions" for personal-level phenomena such as perception (pp. 93–94), but deny that there is any simple one-way relationship from the subpersonal to the personal level, or vice versa. Gallagher and Zahavi specifically reject any "necessary isomorphism between" personal and subpersonal levels (p. 168): a distinction at the subpersonal level need not make its way up to our experience at the

personal level (pp. 168–169), and likewise, phenomenological descriptions of personal-level phenomena "[do] not justify any particular inference to what happens on the subpersonal level" (p. 105). Phenomenological claims at the personal level may indeed be revised because of subpersonal-level discoveries, but this must be due to more careful analyses of our experience made in light of these subpersonal level results, rather than a straightforward inference from the subpersonal level findings. As Zahavi (2010) writes, "Ultimately, the only way to justify a claim concerning…the phenomenological level is by cashing it out in experiential terms" (p. 7).

This discussion is beginning to blend into the methodological issues I will address in chapter 3. What I hope to have shown here, however, is simply that the phenomenological critics seem to adopt a conception of the personal and subpersonal levels close to the one I have sketched here. This will mean my defense of the folk psychological account against their attacks need not deal with a radical difference in how levels of phenomena and explanation in the cognitive sciences are conceived. The phenomenological critics and I are, in a manner of speaking, "playing the same game." My dispute with the phenomenological critics will thus concern particular claims about phenomena at these different levels—in particular, whether the concepts of "theorizing" and "simulation" posited by, respectively, TT and ST make sense as *subpersonal-level* accounts.

Chapter 3. Personal and Subpersonal Level Investigations

1. Personal and Subpersonal Level Inquiries

I will here provide a more comprehensive discussion of how research into social understanding can proceed at personal and subpersonal levels. I will present various investigative techniques used in the cognitive sciences, and the way they provide evidence for phenomena at one or more levels.

In the framework I described in chapter 2, personal-level phenomena are the environmentally situated activities and conscious experiences of human agents, while subpersonal-level phenomena are the operations of the parts of persons, primarily their brains, which themselves are composites made of parts. Mechanistic, constitutive explanations are interlevel accounts of how the organized activities of subpersonal parts enable personal-level phenomena. Personal-level inquiries are thus concerned with describing personal-level phenomena, such as how persons behave in different environmental contexts. Besides behavior, conscious mental activities belong amongst the realm of personal-level phenomena. Subpersonal-level inquiries, in contrast, are concerned with the component parts of persons and their operations. Following Bechtel (2008b), the task of personal-level research can be described as one of *delineating the phenomena* to be constitutively explained in terms of subpersonal processes, and the subpersonal-level task as one of *mechanistic decomposition*, which involves *structural decomposition* (identifying the relevant subpersonal parts) and *functional decomposition* (identifying the operations, i.e., changes or processes

involving the parts). As Bechtel shows, different research techniques are useful for decomposing a mechanism into parts and into operations. Sometimes study of a mechanism's working parts can lead to the realization that the higher-level phenomenon of interest has been misdescribed, requiring revision to how that phenomenon is itself characterized. Given the complexity of the human organism, particularly its brain, it is unsurprising that comprehensive, interlevel explanations are rare at this point in the cognitive sciences.

In this chapter I will address different kinds of research performed at the personal and subpersonal levels. Although the aim is for interlevel constitutive explanations, some techniques more directly target the personal level or a single subpersonal level, while others are more interlevel in nature. I will start in section 2 with the standard, third-personal investigative techniques in the cognitive sciences, as surveyed by Bechtel (2008b). These techniques include behavioral experiments, ethnographic descriptions of behavior and organisms' environments, direct investigations of the brain, and computational modeling. Behavioral experiments and ethnographic descriptions are especially relevant to characterizing personal-level phenomena, but can also provide evidence about the subpersonal-level mechanisms driving these phenomena. The research techniques of neuroscience more directly target subpersonal-level phenomena, but often are inherently interlevel, since the subpersonal phenomena of interest are only those working parts that enable particular personal-level phenomena. An eye to the personal level will thus often be essential when investigating subpersonal-level phenomena. These third-personal techniques do not, however, provide direct evidence of people's conscious experiences, phenomena I

consider crucial to personal-level accounts. While third-person techniques such as behavioral experiments can be designed to provide indirect evidence about our conscious experience, more direct information about consciousness would be preferable. Accordingly, in section 3 I will address issues about acquiring first-person data about our conscious experiences, particularly the role of methods deriving from the phenomenological tradition of philosophy. I will end this chapter by sketching the import this metaphysical and methodological multilevel framework of personal and subpersonal levels for accounts of social understanding, particularly the debate between folk psychological accounts and the phenomenological critics.

2. Third-Personal Research Techniques

2.1. Observing Behavior

2.1.1. Behavioral Studies and the Personal Level. A primary investigative tool in the cognitive sciences is the behavioral experiment. There are several purposes behavioral experiments can serve, but an essential one is differentiating and characterizing personal-level phenomena, i.e., describing how people behave in particular environmental conditions. Real-world behavior is massively complex, so the controlled conditions of behavioral experiments can help to better describe how behavior is affected by stimuli and task conditions of various types. While one must always be concerned about the ecological validity of experimental settings, behavioral experiments are important to determining how the mental mechanism responds to particular stimuli, which cannot be so isolated in real-world settings. The results of

behavioral experiments can be coupled with ethnographic descriptions of real-world environments and behavior in those environments, to help delineate the personal-level phenomena to be given mechanistic explanations.

While I just mentioned ethnography somewhat in passing, it cannot be underestimated how important rich and accurate characterization of environmental structure is to the cognitive sciences. Knowing precisely the environmental conditions in which a mechanism is situated is essential to delineating the phenomena that are enabled by a particular mechanism, and to providing constitutive explanations of how the mechanism's organized parts bring about these phenomena. Chomsky famously argued, for example, that children's environments are too impoverished of linguistic stimuli for them to learn syntactic structure, concluding that syntax must be innate. Part of defending such an argument requires careful study of the language children actually hear during development. Without knowing the actual information children are exposed to in their environments, it is difficult to make any firm conclusions about what is or is not learnable.[12] More generally, the external environment is essential to situated or embedded theories of cognition. These accounts downplay the need for complex subpersonal mechanisms, instead attributing behavioral complexity to the richness of the environment with which organisms interact. While the "information-rich environment, simple mechanism" view is sometimes overstated, adequate recognition of this as a theoretical possibility has come only recently, so it is still

[12] I take this example from Deák, Bartlett, and Jebara (2007).

worth emphasizing. Regardless of the complexity of the subpersonal mechanisms at issue, mechanistic explanations require accurate characterization of the environmental conditions in which a mechanism operates.

It is important to note that since behavioral experiments measure verbal or nonverbal behaviors, rather than the conscious experiences while engaged in these tasks, they usually do not take a stand on this latter feature of the personal level. Behavioral studies thus generally do not tell us to what extent conscious mental processes play a role in driving different types of behavior, or how these conscious experiences relate to subpersonal processes.[13] This is not to say there are no third-person experimental paradigms for investigating conscious versus unconscious mental processes (see, e.g., McGovern & Baars, 2007). For example, the subliminal priming paradigm is used to test the cognitive effects of unconscious perception as compared to conscious perception. As Merikle and Daneman (1998) describe, while early research used first-person, introspective reports to identify when stimuli were not consciously perceived, research performed in the last several decades has used a behavioral measure instead: namely, whether participants can discriminate between alternative stimuli. If participants are unable to discriminate between stimuli, it is assumed that they lack any conscious perception of the stimuli. Researchers can then use such task conditions to detect the cognitive effects of unconscious perception, such as the effects on affective reactions. These kinds of comparisons between conscious

[13] I will discuss more direct, first-person methods for studying consciousness in section 3.

and unconscious cognitive processes have revealed some general characteristics of each, such as the fact that conscious processes seem to be limited in capacity, while unconscious processes are relatively unlimited in capacity (see McGovern & Baars, 2007). Nonetheless, the point remains that most behavioral experiments do not provide data about the extent to which conscious mental processes are driving behavior. As I'll address below in section 3, different kinds of inquiry are required to make such claims.

2.1.2. Behavioral Studies and the Subpersonal Level. Behavioral experiments can also provide indirect evidence of the subpersonal processes enabling personal-level phenomena. As Bechtel (2008b) describes, experiments measuring error rates or reaction times can be used to make inferences about the subpersonal operations responsible for these behaviors. For example, increased reaction time on a behavioral task given one set of stimuli as compared to another may mean the first requires the execution of additional subpersonal operations. While researchers sometimes just set out to record how people respond to a range of stimuli in a given experimental paradigm, behavioral experiments are often constructed so as to test predictions about behavior made by competing models of cognition. For example, researchers have attempted to adjudicate the debate between TT and ST by finding behavioral tasks where the two accounts predict different behavioral responses. Nichols and Stich (1992, 1995; see also Saxe, 2005) suggest that TT and ST differ with regard to the issue of "cognitive penetrability," i.e., sensitivity to false or missing information about the target domain of one's judgment. If we predict and explain people's behavior

through the application of a folk psychological theory, and our theory has gaps, or mischaracterizes relations between variables, this would lead to mistaken predictions and explanations. But if we simulate other people's mental states to predict and explain their behavior, possessing false information or no information about the psychological processes driving people's behavior will not affect our responses. This is because simulation involves the offline use of our own psychological mechanisms for reasoning and decision making—i.e., the same ones by which we and other people reason and make decisions—which are assumed to be cognitively impenetrable and thus unaffected by the theoretical information we have about how these mechanisms of ourselves and others function. The different behavioral predictions of TT and ST can then be tested. One way proposed by Nichols and Stich is to appeal to the surprising behavioral effects identified by social psychologists. If participants are asked how to predict how other people will behave under such conditions, and they make inaccurate predictions, this serves as evidence in favor of TT over ST. ST says we simulate being in those task conditions, use the same mechanisms for prediction we would if we were ourselves performing the task, and thus should make accurate predictions. TT, however, requires that we possess a folk psychological theory covering our behavior in such cases. So folk psychological theorizing will make mistaken predictions without such a theory or with a faulty theory. Predictive errors of this kind thus are claimed to support TT over ST as an account of the psychological processes driving predictions of behavior.

It should be remembered, however, that it is controversial to make definitive conclusions about subpersonal operations from the indirect evidence provided by

behavioral experiments. It is often possible to generate alternative explanations of the cognitive processes driving a particular behavioral effect. In general, predictions generated by one's subpersonal model are preferred to post hoc explanations of how another model can also accommodate the behavioral data. But of course prediction of experimental results is not always required for a subpersonal model to be accepted as the correct explanation of a behavioral study's results. Thus, it is a complicated and contested matter to make inferences about subpersonal operations from behavioral data. This has certainly been the case for the debate between TT and ST. Whatever behavioral competency claimed to be predicted by one theory can often be given an explanation in terms of the alternative account as well. While the number of researchers holding onto a pure TT or ST rather than some form of TT–ST hybrid is dwindling, behavioral evidence has not been very successful at distinguishing accounts of the subpersonal processes driving this aspect of social cognition. Further, it should be noted that the cognitive operations posited from behavioral evidence are usually conceived of in functional terms, without much, if any, regard to their implementation in the brain.

2.1.3. Developmental Studies. One subset of behavioral studies is worth further mention: those in developmental psychology. Many behavioral studies in developmental psychology are intended to provide evidence of already identified personal-level competencies. For example, the "theory of mind" studies described in chapter 1 address the development of various mental state concepts, such as belief. Determining the developmental trajectory of personal-level capacities is a research

problem in itself, addressing questions such as whether such competencies are learned from environmental experience, or mature in everyone regardless of their particular experience. Knowing the developmental trajectory can say something about the subpersonal processes involved, e.g., if we can tease apart simpler from more complex behavioral competencies, showing that they can be dissociated and thus potentially involve different subpersonal processes, or showing how the more mature skill develops out of the simpler one**.**

An aspect of developmental research more relevant to this dissertation (particularly chapter 5), however, is the use of nonverbal tasks to test the cognitive competencies of infants who have yet to develop language. One popular experimental technique to use with nonverbal infants is to measure looking time. For example, children may be habituated to a stimulus of one type (usually defined in terms of a 50% decline in mean looking time across three successive trials compared to the initial three trials), then presented with either a stimulus of the same type, or one of a different type. If children look longer at the "novel" stimulus, it is assumed they are able to discriminate these categories. While there are numerous issues that can be raised about these looking-time studies (see Aslin, 2007), one especially relevant for my purposes is whether nonverbal tasks such as these test the same cognitive competencies as verbal tasks—or more generally, whether different behavioral measures test the same subpersonal-level cognitive capacities. While this is an issue for all behavioral studies, it is especially apparent in developmental contexts, where nonverbal measures are used with young children who have not yet acquired language,

and verbal measures are used with older children possessing more sophisticated language skills.

As Woolley (2006) describes, there is evidence of dissociations in development between verbal and behavioral measures for a wide variety of cognitive domains, including understanding of the distinction between fantasy and reality, and of mathematical equivalence and conservation. Interestingly, the dissociations are in different directions for these two domains of understanding: children show more mature knowledge of the fantasy–reality distinction in their verbal responses than in their nonverbal behavior, but show more mature understanding of mathematical equivalence in their behavioral responses (specifically, their gestures) than in their verbalizations. Accordingly, Woolley offers distinct explanations of the cognitive processes at work in these two cases. In the case of fantasy–reality differentiation, she argues that children possess both fantasy-based and reality-based belief systems, which are equally available to conscious awareness but are differentially expressed in different modalities. In spelling this out, Woolley cites Stubbotsky's (1993) explanation that children hold onto their magical thinking after acquiring more reality-based, scientific beliefs, but with experience learn that adults value the scientific world view, so tailor their verbal responses to meet these societal explanation. When freed from these societal constraints, children will continue to express these beliefs in their nonverbal behavior. In contrast, in the case of mathematical equivalence and conservation, Woolley argues that nonverbal responses express a nascent implicit understanding that is not available to consciousness, while verbal responses require

the development of explicit knowledge before they can achieve successful responses in this modality.

While I have only sketched Woolley's account, it illustrates some of the complexities of developmental research, with regard to the behavioral phenomena to be explained and the subpersonal-level processes posited to explain these phenomena. There are important conceptual and methodological questions about how to determine when two observable behaviors express the same underlying cognitive capacity. I will be addressing this issue somewhat in chapter 5 when I discuss the phenomenological critics' claim that online social understanding is driven by cognitive processes distinct from those driving offline social understanding.

2.2. Observing the Brain

While behavioral studies can provide us with some general ideas of the subpersonal operations enabling these personal-level behaviors, investigating the brain provides more direct evidence of the parts and the operations performed by these parts needed for mechanistic, constitutive explanations of personal-level phenomena. One subpersonal task is that of structural decomposition: to differentiate the various parts of the brain, and decompose these parts into their parts, and so on. Since mechanistic explanations of particular personal-level phenomena are concerned with not just static structures but working parts, mechanistic decomposition must be performed in a task-specific manner, with an eye to the components whose operations are involved in enabling a particular phenomenon. Sticking to the subpersonal level, one way to study the operations of parts is to intervene on one part and see its effect on other parts,

measured via techniques such as single cell recording, EEG, PET, and fMRI. *Lesion experiments* are one form of subpersonal intervention, where a part is temporarily or permanently disabled, whereas *stimulation experiments* involve stimulating the activity of a part.

Given the hierarchy of subpersonal levels, there is much work to be done at these various levels. But it is unlikely that much can be learned for the purpose of mechanistic explanations of personal-level phenomena with a purely subpersonal approach. Most brain research takes an *interlevel* approach, keeping in mind the personal-level phenomena of interest when examining brain activity (see, e.g., Craver, 2007). Taking a bottom-up approach, the role of particular parts in producing a personal-level effect can be investigated via lesion or stimulation studies that measure the effect on behavior of these subpersonal manipulations. Taking instead a top-down approach, neuroimaging or other techniques for measuring brain activity may also be used to determine the brain areas that are relatively more or less active while performing different behavioral tasks. Careful attention to the kinds of tasks for which a brain area is responsible can help to determine the operations performed by this part. There is a concern here, however, that neuroimaging studies simply identify the brain areas which are active during particular behavioral tasks, without being able to say anything about what operations these parts are performing. Or worse, a single brain part may be attributed full responsibility for the operations enabling some behavioral effect. This can happen when researchers fail to remember that many neuroimaging studies use a subtractive method, comparing brain activity across two different tasks (see Bechtel, 2008b, p. 47, in press, §4). One of these tasks is thought to employ fewer

cognitive operations compared to the other, so the brain activity observed during the simpler task is subtracted away from the brain activity observed during the more complex task, leaving the (increased or decreased) activity of brain areas thought to be unique to the complex task. The problem comes when these brain areas are assumed to be solely responsible for the cognitive operation(s) unique to the complex task, forgetting the contribution of the subtracted areas and ignoring potential interactions between areas which might alter the operation of these areas across the two tasks. While there are certainly difficulties in identifying the operations performed by particular brain parts, it is an essential component of multilevel, mechanistic explanations in the cognitive sciences.

2.3. Computational Modeling

Another major methodology in cognitive science for determining the operations performed by subpersonal mechanisms is computational modeling. Given our limited understanding of the brain, most computational models in cognitive science function as what Craver (2007) calls "how-possibly models" of cognitive mechanisms: they are conjectures about possible component parts, the operations they might perform, and how they might be organized so as to together perform the cognitive phenomena of interest (see also Bechtel & Abrahamsen, in press). The main empirical constraint on computational models in cognitive science is how well they approximate the personal-level behaviors of whole human beings. We currently lack the neuroscientific understanding necessary in most cases to confirm whether the mechanisms posited by computational models in any way fit our actual cognitive

mechanisms. As Bechtel (2005, 2008a; Bechtel & Abrahamsen, in press) notes, it is not clear that the two major classes of computational models used in cognitive science research—symbol-processing and connectionist models—are adequate characterizations of our actual cognitive mechanisms. Thus, computational models are primarily used to construct models of what our cognitive mechanisms might be like, and experiment on the capacities and limitations of such systems compared to those of actual organisms.

A new trend in cognitive modeling is what Deák et al. (2007) call "developmental systems modeling." This approach ambitiously attempts to model artificial embodied agents with biologically plausible neural systems engaging in a history of sensorimotor interaction with complex physical and social environments. Such models require detailed ethnographic data about agents' environments, as well as neuroscientific data about the relevant neural mechanisms, such as that precise formal models can be constructed of how these subpersonal mechanisms might operate and change over time. If the behavior of artificial agents in computational models matchs data about the behavior patterns of real people, we have reason to believe the model might have the right story about the subpersonal mechanisms driving these behaviors. The anchoring of computational models in empirical data about the components and activities of cognitive mechanisms and the environments in which they are situated is an important trend that should continue if we are to develop more complete, accurate mechanistic explanations of cognition.

3. Phenomenological, First-Person Methods, and their Relation to Third-Person Data

The third-person perspective of behavioral studies limits them to claims about the environmental conditions and verbal and nonverbal behaviors of human agents. They thus are unable to say much about what I have identified as another essential feature of personal-level accounts: people's conscious experiences. Neuroscientific methods for observing the brain similarly fail to provide direct evidence about conscious experience. So how we are to obtain reliable data about consciousness? Given the limitations of third-person methods, first-person methods—i.e., methods where subjects describe their own experiences—seem necessary. But concerns about the scientific rigor of first-person reports have historically been so great as to lead to the behaviorist movement in psychology, which eschewed all references to the mind. Issues with regard to introspective reports include: How accurately do verbal reports (i.e., linguistic expressions of our beliefs about our experiences) characterize our actual experiences? What methods are required to obtain reliable, consistent, and valid introspective reports? How do introspective reports relate to third-person characterizations of mental processes? Only in the last decade or two has consciousness returned as a reputable scientific subject matter where such questions are asked and answers are attempted (see, e.g., Jack & Roepstorff, 2003; Roepstorff & Jack, 2004).

3.1. Phenomenological Method of Describing Conscious Experience

Given that my focus in this dissertation is on criticisms of folk psychology offered by philosophers from the phenomenological tradition of Husserl, Heidegger, Merleau-Ponty, etc., I will focus on first-person methods derived from this tradition. A unique feature of phenomenology is that its adherents draw a distinction between it and introspective forms of psychology. Introspection is characterized as a kind of internal perception directed towards our minds, an awareness of our "inner" mental states. Phenomenology, in contrast, is best understood as a transcendental philosophical project in the tradition of Kant, attempting to characterize the way the world appears to creatures like us and the conditions of possibility for this experience (Gallagher & Zahavi, 2008, ch. 2).

While there is much that could be said about the philosophical, transcendental project of phenomenology as originated by Husserl and developed by later philosophers, there are at least two elements that distinguish phenomenology from introspection. First, since phenomenology is concerned with the way the *world* appears to us, it is not restricted to describing an "inner" mental realm, which is how introspection is commonly characterized. Phenomenological reports thus concern the *objects* experienced as much as they concern the subjective side of consciousness. Note that this distinction is wholly agnostic about subpersonal-level mechanisms. It is a distinction in the content of the reports produced, in what the subject focuses on in producing these reports, rather than in any sort of mechanistic explanation of the psychological processes involved in phenomenological reflection versus introspection. This is tied to the transcendental, philosophical aims of phenomenology—phenomenology aims to do more than introspective psychology by describing the

conditions of the possibility of our experiences. This leads in to the second element I want to highlight: phenomenologists attempt to describe intersubjectively accessible, invariant structures of our experience of the world, rather than merely the subjective, qualitative features of my individual experience, as introspection is often characterized as doing. In Gallagher and Zahavi's (2008) words:

> Phenomenology is not interested in understanding the world according to Gallagher, or the world according to Zahavi, or the world according to you; it's interested in understanding *how it is possible* for *anyone* to experience a world. In this sense, phenomenology is not interested in qualia in the sense of purely individual data that are incorrigible, ineffable, and incomparable.… Phenomenology is interested in the very possibility and structure of phenomenality; it seeks to explore its essential structures and conditions of possibility. (p. 26)

By taking a transcendental perspective, phenomenology treats consciousness in a very different way compared to introspection—not simply as what is going on in my mind, but as how the world is being disclosed to, or "constituted" for, a conscious subject. Gallagher and Zahavi (2008) summarize the transcendental philosophical project of phenomenology in the following "somewhat paradoxical way":

> …phenomenologists are not interested in consciousness per se. They are interested in consciousness because they consider consciousness to be our only access to the world. They are interested in consciousness because it is world-disclosing. Phenomenology should therefore be understood as a philosophical analysis of the different types of world-disclosure (perceptual, imaginative, recollective, etc.), and in connection with this as a reflective investigation of those structures of experience and understanding that permit different types of beings to show themselves as what they are. (p. 26)

The transcendental aims of phenomenology come through when the phenomenological method is described in more detail. As Gallagher and Zahavi

(2008) present it, the phenomenological method consists of four main steps, which I will list then unpack below:

1. The *epoché* or suspension of the natural attitude.
2. The *phenomenological reduction*, which attends to the correlation between the object of experience and the experience itself.
3. *Eidetic variation*, which keys in on the essential or invariant aspects of this correlation.
4. *Intersubjective corroboration*, which is concerned with replication and the degree to which the discovered structures are universal or at least sharable. (p. 29)

The first two steps concern the stepping back from one's beliefs and commitments about experience, so as to most accurately describe experience itself. The "natural attitude" is Husserl's term for the naïve realism found both in the sciences and in our daily lives: we assume the existence of a mind-independent world, which we attempt to discover and investigate through everyday experience and the refined methods of science. Phenomenologists believe we must "bracket" or "suspend" such assumptions about the world, so as to consider the way the world appears to the subject, and the contributions of the subject in this "constitution" of the world of experience. This attitude of suspending our pre-theoretical, realist commitments is called the *epoché*. As Overgaard, Gallagher, and Ramsøy (2008) summarize:

> In effect, phenomenology does not appeal to scientific or metaphysical explanations of the world, or our experience of it, nor is it looking for an analysis cast in terms of common sense or folk psychology. By clearing

> away our ordinary opinions, our everyday attitudes about things, and even our scientific theories about how things work, the aim is to get at the world as it is experienced, and in particular to describe how things appear in that experience. (p. 105)

Tied to the epoché is the *phenomenological reduction*, whereby we focus on the subjective contributions to our experience of the world, the invariant structures by which an experience of a particular object is made possible. For example, we focus not only on the object perceived, say an apple, but on what it is that makes this experience one of perception, as opposed to memory or imagination. In this way we attend to the "correlation" between the object perceived and the experience of perception.

The general shift in attitude toward our experiences provided by the epoché and phenomenological reduction clearly expresses the transcendental aims of phenomenology. Given this attitude, *eidetic variation* characterizes one approach to obtaining phenomenological descriptions. It involves changing or removing various features of the experience, so as to determine what is essential or invariant to experiences of different types. For example, as Thompson, Noë, and Pessoa (1999) explain, Merleau-Ponty (1945/2002) uses this method to identify the figure–ground structure as an invariant feature of perception:

> no matter how one imagines the perceptual situation to be varied, the figure–ground structure always remains as a formal, constitutive feature of perception, while on the other hand, imagining the figure–ground structure to be absent is tantamount to no longer imagining a case of *perception*. (Thompson et al., 1999, p. 189)

Admitting these descriptions are defeasible leads naturally to the fourth step of *intersubjective corroboration*, comparing one's descriptions against others, and going

through the messy process of determining how to best characterize experience in general.

Independent of the transcendental philosophical commitments of phenomenology, the core idea behind the phenomenological method is to "bracket" one's beliefs about the nature of experience so as to describe as accurately as possible experience itself. Everyone should recognize this as a goal for first-person data. Whether specific training is required to meet this goal, as the phenomenologists claim, is an open question in the science of consciousness studies. I take no stand here on whether phenomenological methods are the best approach to first-person investigations of conscious experience. I have focused on phenomenological methods rather than other first-person methods to better understand the methodological background for the claims of the phenomenological critics of folk psychology. It is worth raising the issue, however, of whether the transcendental approach of phenomenology is compatible with the naturalism of cognitive science. If phenomenology is a form of transcendental philosophy, can phenomenological descriptions of consciousness be treated like other forms of evidence in the cognitive sciences? Can they be treated on par with other first-person method of describing conscious experience? Or are they different because they are aiming at the conditions of the possibility of experience? Exactly what a "naturalized phenomenology" would look like is a topic of debate amongst phenomenologists (e.g., Roy, Petitot, Pachoud, & Varela, 1999; Zahavi, 2004b, 2010). The view of the phenomenological critics of folk psychology (Gallagher and Zahavi in particular) is that although phenomenology is a philosophical rather than empirical or scientific enterprise, phenomenological

descriptions of consciousness certainly can be of use to empirical work in the cognitive sciences. There are numerous proposals about how phenomenology and the cognitive sciences might be related, some of which I'll address in the next section.

3.2. Relating First-Person and Third-Person Data

So how are first-person data obtained by phenomenological methods supposed to be related to third-person research methods in obtaining mechanistic, constitutive explanations of personal-level phenomena? Gallagher and Zahavi (2008, ch. 2) summarize three proposals for integrating phenomenological methods with other, third-person research techniques:

1. *Mathematical formalization* (Roy et al., 1999; Yoshimi, 2007): Formalize phenomenological descriptions using the mathematics of dynamical systems theory, such that these mathematical descriptions of first-person experience can be related to mathematical characterizations of third-person data about subpersonal-level phenomena. Mathematics is seen as a neutral realm to integrate first- and third-person data, to construct integrated multilevel mechanistic explanations. Dynamical systems theory in particular is emphasized as a tool for characterizing phenomenological data "insofar as conscious processes unfold over time in a structured way" (Yoshimi, 2007, p. 286). Dynamical systems theory and (Husserlian) phenomenology "are both founded on the basic principle that systems can be understood in terms of their possibilities. One finds in each case a space of possibilities, which is multi-dimensional, has topological and geometric structure, and whose members

must be instantiated in accordance with rules in order for coherent behavior to arise" (Yoshimi, 2007, p. 272).

2. *Neurophenomenology* (Lutz & Thompson, 2003; Varela, 1996): Train participants in phenomenological methods, obtain first-person descriptions of their experience (using their own, intersubjectively validated theoretical categories), while also measuring brain activity (using, e.g., EEG). This permits connections to be drawn between the dynamics of conscious experience and the dynamics of neural processes.
3. *Front-loaded phenomenology* (Gallagher, 2003): Use concepts and distinctions from phenomenological analyses to design experiments, which do not necessarily require phenomenological training or even first-person reports from participants. For example, Gallagher (2003) describes neuroimaging experiments (e.g., Farrer & Frith, 2002) that appeal to a phenomenological distinction in our experience of action: the sense of *ownership* (the sense that it is *my* body that is moving rather than someone else's) versus the sense of *agency* (the sense that I intended or caused the movement, rather than some external force being the cause of my movement). While the senses of ownership and agency coincide and are indistinguishable in normal cases of action, experimental tasks can be designed to dissociate them, so their neural correlates can be studied via neuroimaging techniques. For instance, Farrer and Frith's (2002) subjects manipulated a joystick to move an image on a computer screen while in an fMRI scanner. In some trials, the image was indeed controlled by the subject's movement of the joystick. But in others, the

experimenter controlled the image's movement, so the subject's movement of the joystick only tracked the movement of the image without causing this movement. These two task conditions kept the sense of ownership constant (since subjects were indeed moving the joystick in both conditions) but varied whether the subjects' sense of agency, allowing the experimenters to explore the neural correlates of the sense of agency. As can be seen in this example, in Gallagher's method of front-loaded phenomenology, rather than providing empirical data directly, phenomenology here serves the same role as any other psychological theory or folk intuitions about the personal-level phenomena to be studied experimentally.

More should be said about how the first-person data obtained via phenomenological methods is to be related to the data obtained via third-person methods. Neurophenomenology suggests third-person methods are inadequate to obtain rich data about our conscious experiences, and thus must be supplemented by phenomenological methods. Phenomenology in this way fills in the evidential gaps left by third-person methods. Front-loaded phenomenology offers a more complex relation between first-person and third-person data. The first-person data are here used to construct experiments designed to obtain third-person data about personal-level phenomena (behavior) and subpersonal-level phenomena (brain activity)—although first-person data could also be obtained in these experimental conditions as well. Despite these differences, neurophenomenology and front-loaded phenomenology both involve a one-way relation between phenomenological methods and third-person

methods—i.e., they describe ways that phenomenology can supplement and modify third-person methods.

But what about the other direction of influence? Do phenomenological methods provide incorrigible data about our conscious experiences, or can they be informed by data obtained by third-person methods? As mentioned at the end of chapter 2, some phenomenologists see the potential for two-way interaction between phenomenological, first-person data and third-person data. Appealing to resources from both classical and contemporary phenomenology, Zahavi (2009, 2010) argues that third-person data from the cognitive sciences can lead us to reexamine and potentially revise our phenomenological descriptions. He gives the following example of how this might work:

> Let us assume that our initial phenomenological description presents us with what appears to be a simple and unified phenomenon. When studying the neural correlates of this phenomenon, we discover that two quite distinct mechanisms are involved; mechanisms that are normally correlated with distinctive experiential phenomena, say, perception and memory. This discovery might motivate us to return to our initial phenomenological description in order to see whether the phenomenon in question is indeed as simple as we thought. Perhaps a more careful analysis will reveal that it harbors a concealed complexity. (Obviously, one might also consider the reverse case, where the phenomenological analysis presents us with what appears to be two distinct phenomena and where subsequent neuroscientific findings suggest a striking overlap, unity, or even identity). However, it is very important to emphasize that the discovery of a significant complexity on the sub-personal level—to stick to this simple example—cannot by itself force us to refine or revise our phenomenological description. It can only serve as motivation for further inquiry. There is no straightforward isomorphism between the sub-personal and personal level, and ultimately the only way to justify a claim concerning a complexity on the phenomenological level is by cashing it out in experiential terms. (Zahavi, 2009, p. 166; see also Zahavi, 2010)

Therefore, while phenomenologists see consciousness as a metaphysically distinct type of personal-level phenomenon that can only be directly accessed through first-person methods, they do accept that empirical data about subpersonal-level phenomena obtained through third-person methods can lead phenomenologists to reexamine their descriptions of conscious experience.

In conclusion, it is increasingly being recognized that first-person methods should be added to the toolkit of cognitive science so as to provide more direct evidence of people's conscious experience. Although phenomenology is a philosophical enterprise with uniquely philosophical aims, including characterizing the conditions of the possibility of experience, it is increasingly recognized that phenomenological methods can be used to obtain descriptions of conscious experience. These phenomenological descriptions can be used as first-person data about some of the personal-level phenomena to be explained in the cognitive sciences. But since they are not treated as incorrigible, phenomenological descriptions are up for revision in light of discoveries from third-person methods.

4. Import for Accounts of Social Understanding

So how does the mechanistic framework developed in this and the previous chapter help to frame the issues raised by the critics of folk psychology about the nature of human social understanding?

First, the distinction between reflective, offline and unreflective, online social understanding should be seen as relevant to *delineating the phenomena* of human

social understanding. While the folk psychological tradition has emphasized reflective, offline social understanding, there exist a vast range of personal-level phenomena involving online social interaction which have not been adequately described and explained. Accounts of social understanding must address these phenomena as well. Behavioral experiments, ethnographic descriptions, and phenomenological, first-person inquiries are required to characterize the personal-level phenomena at issue here. The next two chapters, but especially chapter 5, address unreflective, online phenomena not emphasized in traditional folk psychological accounts: namely, phenomenologically-direct social perception of other people's intentions and emotions, and online social interactions which do not involve conscious belief–desire attribution.

Further, there are questions of whether the folk psychological tradition has adequately characterized the nature of reflective, offline social understanding. Do folk psychological attributions of mental states exhaust the kinds of offline explanations and predictions of behavior humans actually give? The debate between TT and ST has kept our focus somewhat narrowly on mentalistic understanding, and failed to recognize the importance of offline social cognition which does not involve the attribution of mental states. While this criticism treats the folk psychological picture as overly narrow, the picture of offline folk psychological reasoning itself has been criticized. For example, offline folk psychological reasoning has traditionally been characterized as something done by individuals from a "spectatorial" stance toward "third-persons." But critics such as Dan Hutto (2004, 2008a) contend that reflective, offline folk psychological reasoning is often, if not primarily, a social practice

engaged in by "second-persons." We engage in discourse with people about their reasons, our own, and those of other people. This social dimension to offline folk psychology has been given little attention in the literature, but must be included in the phenomenal description of human social understanding.

Issues of delineating personal-level phenomena naturally lead to issues concerning the *mechanistic decomposition* of human social understanding. Here is where we are confronted with the criticism that the folk psychological reasoning characterized by TT and ST is not a pervasive feature of human social understanding because it does not underlie phenomenologically-direct social perception or online social interaction. Although these critics do not always do so, this criticism should be focused on TT and ST as *subpersonal-level* accounts, since they are accepted as *personal-level* accounts of some of our conscious psychological processes involved in offline explanation and prediction of behavior. The proposal of the phenomenological critics is thus that subpersonal theorizing and simulation of people's mental states is not what constitutively explains direct social perception and online social interaction. They thus suggest we must look to subpersonal-level accounts other than TT and ST when offering constitutive explanations of these personal-level phenomena.

I accept that characterizations of the phenomena of human social understanding and mechanistic explanations of these phenomena must become more pluralistic—that there are phenomena at both personal and subpersonal levels which are not adequately captured by the traditional folk psychological picture. The central issue I am addressing in this dissertation, however, is how pervasive folk psychological reasoning is to human social understanding. Are the personal-level

phenomena identified by the phenomenological critics—i.e., direct social perception and online social understanding—driven by different subpersonal-level processes than those enabling offline reflection on people's behavior? What kind of evidence is needed to answer this question?

We must remember here that subpersonal findings can bring revision to our characterization of personal-level phenomena. Finding that the same mechanisms are operative in personal-level tasks we have characterized by different names suggests these are closer in nature than we originally assumed. Of course, social understanding is too complex to be a matter of whether a single mechanism enables all or only some of the personal-level phenomena of interest. So the issue here will really be whether the same *kind* of mechanistic processes drive both the offline phenomena emphasized by the traditional folk psychological accounts, and the phenomena emphasized by the phenomenological critics—namely, whether the subpersonal mechanisms for the phenomena at issue make use of folk psychological theorizing and simulation, or mental state attribution simpliciter. I will argue in the next two chapters that the phenomenological critics have given us no convincing reasons to think so, and that theorizing and simulation remain viable subpersonal-level accounts. In chapter 4 I will argue that this is the case with regard to direct social perception, defending the viability of subpersonal versions of TT and ST as explanations of this class of personal-level phenomena identified by the phenomenological critics. In chapter 5 I defend the proposal that subpersonal-level attributions of beliefs and desires may drive our online social interactions.

Chapter 4. Direct Social Perception: A Challenge to Theory Theory and Simulation Theory?

1. Folk Psychological and Phenomenological Accounts of Social Perception

A core assumption of folk psychological accounts is that our perceptual experience of other people can only be of their physical movements, since mental states are unobservable. To appreciate people's mental states, an extra psychological step beyond perception is needed. The two major folk psychological accounts, TT and ST, differ with regard to the kind of psychological process used to go from perceptual information about observable behavior to the attribution of a mental state. Mental state attribution occurs for TT via theoretical inference, by applying theoretical knowledge about the relations between observable behavior, environmental context, and mental states. ST denies that we possess such theoretical knowledge, instead claiming that we simulate being another person, determine what mental states we would have if we were that person in that situation, and project those mental states onto the other person. For both folk psychological accounts, our perceptual experience is of mere behavior, and mental state understanding only comes when we perform some extra psychological process beyond perception.

As introduced in chapter 1, this folk psychological picture of social perception has recently come under attack by the phenomenological critics, who claim that we "directly perceive" some mental states (Gallagher, 2001, 2005, 2007, 2008b; Gallagher & Zahavi, 2008; Ratcliffe, 2007; Zahavi, 2005, 2007, 2008). Zahavi (2005),

for example, writes: "…the life of the mind of others is visible in their expressive behavior and meaningful action" (p. 222). Scheler and Wittgenstein are often identified as historical proponents of this view, with passages such as the following:

> For we certainly believe ourselves to be directly acquainted with another person's joy in his laughter, with his sorrow and pain in his tears, with his shame in his blushing, with his entreaty in his outstretched hands, with his love in his look of affection, with his rage in the gnashing of his teeth, with his threats in the clenching of his fist, and with the tenor of his thoughts in the sound of his words. If anyone tells me that this is not "perception," for it cannot be so, in view of the fact that a perception is simply a "complex of physical sensations," and that there is certainly no sensation of another person's mind nor any stimulus from such a source, I would beg him to turn aside from such questionable theories and address himself to the phenomenological facts. (Scheler, 1954, p. 260)

> ''We see emotion."—As opposed to what?—We do not see facial contortions and *make the inference* that he is feeling joy, grief, boredom. We describe a face immediately as sad, radiant, bored, even when we are unable to give any other description of the features.—Grief, one would like to say, is personified in the face. (Wittgenstein, 1980, vol. 2, §570)

These two passages concern the perception of others' emotions. Intentions have also been explicitly identified as directly perceivable. Although this is not always noted, this claim concerns intentions that are concurrent with the performance of an action ("intentions in action"), rather than intentions to act at some future time ("prior intentions). But as far as I know, no one has argued that beliefs are perceivable in his way. Most likely desires will fall with beliefs in the category of unperceivable states—Gallagher (2001), for instance, identifies both beliefs and desires as "hidden away" in people's minds (p. 86). Accordingly, I will here be concerned with the claim that some subset of mental states, particularly *intentions* and *emotions*, are directly perceivable.

Authors differ in their descriptions of this experience of other persons, calling

it either a form of "direct perception" (Gallagher, 2007, 2008a) or a "distinctive mode of consciousness, different from perception, recollection and fantasy" called "empathy" (Zahavi, 2007, p. 36; see also Thompson, 2001). But all those in this camp agree that we can "directly" experience—i.e., without a mediating conscious psychological process, such as inference—other people's mental lives. Further, they claim that direct social perception is more pervasive in our everyday social lives than the conscious reflection emphasized in folk psychological accounts.

As Gallagher (2008b) notes, the idea of "direct perception" is often associated with Gibson (1979). By "direct perception," Gibson meant that all the information we need to perceive objects or possibilities for action is already in the light transmitted to our sense organs, ready to be "directly" detected by us; it need not be inferred or computed from sensory stimulation via additional psychological processes. What exactly Gibson meant by "directly" is a matter of dispute. One interpretation mirrors the phenomenologists' sense of direct perception, reading "no additional psychological processes" as meaning no additional *conscious* psychological processes. But even granting this, no one can deny that perception involves complex processes inside a person, particularly in the brain. In other words, even given the directness of perception at the personal level (i.e., the level of conscious experience), there remains a story to be told at the subpersonal level of the brain processes (and potentially other non-neural internal processes) enabling perception. Gibson is also often read as rejecting information-processing or representational characterizations of these subpersonal processes. Whatever Gibson's views on the subject, it is a central issue for cognitive science to characterize the subpersonal-level processes enabling personal-

level psychological phenomena such as perception.

Focusing on this distinction between personal- and subpersonal-level accounts, I will here evaluate Dan Zahavi and Shaun Gallagher's respective objections to TT and ST as accounts of direct social perception. I will argue that their phenomenology-based criticisms are much more narrowly focused than they appear. While they do have bite against *personal level* versions of TT and ST, they do not rule out these theories as accounts of the *subpersonal level* processes enabling direct social perception. Further, I show that their arguments directly addressing the subpersonal level—particularly Gallagher's rejection of a subpersonal-level notion of simulation—are unconvincing. To be clear, in this chapter I remain agnostic about TT and ST as accounts of the subpersonal processes underlying direct social perception. Indeed, I leave open whether *any* subpersonal processes should be characterized in these terms. My aim is instead to distinguish personal- and subpersonal-level accounts of direct social perception and expose the limitations of Zahavi and Gallagher's criticisms about subpersonal-level versions of TT and ST.

2. Zahavi Against Theory Theory

In his book *Subjectivity and Selfhood*, Dan Zahavi (2005, ch. 7) presents arguments against TT's account of self- and other-awareness. Although the two are related, I will focus on Zahavi's arguments against TT's account of other-awareness—specifically, our ability to attribute mental states to others, which Zahavi here refers to as "mindreading." Zahavi's challenge to TT focuses on the connections it draws

between mindreading, possessing a theory of mind, and passing false-belief tasks. It can be summarized as follows (2005, pp. 197, 214):

1. TT proposes that to mindread, one must possess a theory of mind.
2. According to TT, passing false-belief tasks is necessary and sufficient for possessing a theory of mind.
3. Children do not pass false-belief tasks until around age four.
4. Therefore, according to TT, children do not possess a theory of mind until around age four.
5. Therefore, TT is committed to the claim that children cannot understand others' mental states until around age four.
6. But children can perceive others' emotions and intentions prior to age four (for evidence from developmental psychology, see Zahavi, 2005, pp. 206–214).
7. Thus, theory theorists must either: (a) adopt an inclusive definition of mindreading and admit that children can mindread before they can pass theory of mind tasks (i.e., false-belief tasks); or (b) retain an exclusive definition of mindreading in terms of false-belief understanding, but concede that children acquire an understanding of emotions and intentions prior to being able to mindread (understood as theorizing about the mind).

The force of the dilemma in (7) is supposed to be that, either way, theory theorists must admit that at least some mental states are understood non-theoretically. Unfortunately, several aspects of this argument are problematic. I will not address evidence against (3), such as Onishi and Baillargeon's (2005) purported evidence of false-belief understanding in 15-month-olds, which potentially collapses the

developmental gap between false-belief understanding and the understanding of emotions and intentions.[14] What I will focus here on is premise (2), and its role in the dilemma Zahavi poses in (7). Most researchers, including theory theorists, now accept that undue attention has been paid to false-belief understanding (see, e.g., Bloom & German, 2000). Accordingly, most theory theorists now reject (2). While false-belief understanding is sufficient for possessing a fairly mature theory of mind, it is no longer treated as necessary; understanding of mental states other than belief is treated as evidence for possessing a theory of mind. In other words, theory of mind is now treated as multifaceted, with some aspects developing prior to false-belief understanding.

Given its dependence on (2), the dilemma in (7) is thus problematic. While the two horns differ in what falls under the heading of mindreading, both depend on (2), which identifies tests of theory of mind with tests of false-belief understanding. As noted above, however, theory theorists no longer treat false-belief understanding as criterial for possessing a theory of mind. Therefore, the force of Zahavi's dilemma for TT is undermined.

There is, however, a substantive issue that emerges from Zahavi's (2005) discussion, one which he himself takes up on the last pages of his book (pp. 221–222). Zahavi admits that we do sometimes consciously theorize about people's mental states, and, following Frith (2003), suggests that this is how some high-functioning

[14] While not focusing on the developmental question, I will discuss this experiment and other similar studies of false-belief understanding in chapter 5.

autistic people are able to overcome some of their deficits in social understanding. The evidence Zahavi provides in favor of (6)—about children's understanding of others' emotions and intentions—does not, however, involve such conscious, reflective understanding. For example, Zahavi (2005, p. 212) describes the capacity for social referencing found in children during their second year of life. When in an unfamiliar situation, infants look to their parent's faces to gauge their emotional reactions to the situation. Infants then use this information about the status of the environment to guide their own actions. If a parent expresses, say, fear toward an unfamiliar object, the child will recognize this negative reaction toward the object and, accordingly, avoid it. It is unreasonable to characterize infants as consciously reasoning from facial expressions to emotional reactions; further, it does not fit the phenomenological experience of adults, as characterized by the quotations from Scheler and Wittgenstein above in section 1. As a result, Zahavi (2005) describes such instances of social perception as "immediate, pre-reflective, or implicit understanding" of others' mental states (p. 221).

Therefore, we can phenomenologically distinguish two capacities: (a) the ability to consciously theorize about people's mental states; and (b) the ability to directly, i.e., non-inferentially, perceive their mental states. Allowing that TT may explain the former, the key question then concerns whether our capacity for phenomenologically direct social perception is driven by *subpersonal* theorizing of which we are not conscious. Zahavi's clearly thinks it is not. But what is the evidence for this conclusion?

Unfortunately Zahavi does not offer much of a defense. He describes Baron-Cohen's (1995) proposed "intentionality detector" and "eye-direction detector" as providing "direct and non-theoretical understanding" of people's intentions and perceptions (Zahavi, 2005, p. 214). But Zahavi offers nothing to support the claim that such mechanisms are in fact non-theoretical. This also comes up in Zahavi's discussion of autism, mentioned briefly above. Uta Frith (2003) claims that autistics lack a "theory of mind module that allows for an intuitive and automatic attribution of mental states to others," but that autistics "might acquire a conscious theory of mind by way of compensation" for this lack (Zahavi, 2005, p. 222). Zahavi accepts that autistics may consciously and explicitly theorize about people in order to make up for their social deficits, but rejects Frith's characterization of nonautistic social perception as theoretical. Yet he does no more than assert this claim, writing that it is "better to avoid using the term 'theory' when speaking of a nonconscious information-processing mechanism" involved in social perception, and that he finds "it rather misleading to designate such nonconscious inferential processes as intuitive" (p. 222).

I do not want to defend the idea that autistics' compensatory, conscious theorizing about people's minds should be equated with nonautistic social perception. Surely there are differences between these phenomena. My point is that the phenomenological considerations offered by Zahavi are insufficient to rule out TT as an account of ordinary social perception. Phenomenological claims simply do not have purchase with regard to nonconscious, subpersonal processes, as Gallagher (2005, p. 215) admits. Theory theorists do usually characterize our social experience as of mere behavior, from which we infer people's mental states. Yet as a claim about

subpersonal-level processes, it seems TT could accommodate the idea that our personal-level, phenomenological experience is not of mere behavior, but of "expressive behavior," i.e., behavior expressive of people's mental states (Zahavi, 2007, 2008). Perhaps TT needs an enriched account of our mental state concepts to accommodate this fact. It looks plausible, however, that the subpersonal-level processes producing such experience might be theoretical in nature. Phenomenological evidence will not speak to this possibility.

So what is it that grounds Zahavi's rejection of TT at the subpersonal level? At the beginning of his chapter on this topic, Zahavi (2005, p. 181) mentions Blackburn's (1992) "promiscuity objection" to TT: the concern that characterizing subpersonal processes as "theoretical" would make this concept entirely vacuous, so almost any belief-formation process would count as theoretical. Zahavi seems persuaded by this worry, such that he treats theorizing and the related concepts of explanation and prediction as exclusively personal-level concepts characterizing conscious, reflective phenomena. Gallagher (2005) explicitly asserts this claim, writing: "Explanation (or theory) seems to mean (even in our everyday psychology) a process that involves reflective consciousness" (p. 215). Such a strict conception of theorizing rules out the possibility that (fully) subpersonal processes—such as those enabling direct social perception—are theoretical in nature.[15]

[15] Since even conscious theorizing must ultimately be characterized subpersonally, the idea must be that the subpersonal processes underlying conscious theorizing are very different from the subpersonal processes enabling direct social perception, such that "theorizing" is never an appropriate characterization of the latter.

I acknowledge Zahavi's concern with trivializing the concept of theorizing. Both proponents and critics of TT have often been quite permissive in the use of this concept. For example, Shaun Nichols and Stephen Stich (2003) use "theory theory" to refer to any "information-rich" process of mental state attribution, i.e., any process which is guided by "a rich set of mental representations containing substantial information (or, sometimes, *mis*information) about mental states and their interactions with environmental stimuli, with behavior, and with each other" (p. 102). Simulation processes, which do not require such bodies of information, are cases of "information-poor" processes. Such a characterization of TT would indeed allow many different kinds of subpersonal psychological processes to be characterized as theoretical. For example, this definition of TT does not mark the distinction between "classical" cognitive architectures (i.e., "rules and symbols" approaches sometimes derided with the label "good old fashioned artificial intelligence") and connectionist architectures. Some have thought true theorizing to require the rule-based manipulation of symbolic representations, and thus rejected connectionist networks as vehicles of theorizing. Moving from features of representational vehicles to issues of representational content, this permissive characterization of TT covers any information about mental states, environmental conditions, and behavior. It does not matter whether this information is a unified, coherent, abstract set of laws, or a less cohesive collection of statistical patterns, algorithms and heuristics. It has been argued that only the former and not the latter deserve to be called "theoretical" in nature.

But even those who use the term "theory" rather permissively admit such distinctions between types of subpersonal processes and draw stronger or weaker

connections between these various phenomena and the conscious theorizing of adults (scientists in particular). Gopnik and Meltzoff (1997), for example, quite specifically define theories in terms of their structural features (abstract, coherent representations of the causal structure of some target domain), functional features (their use for explanation, prediction, and interpretation) and dynamic features (their defeasibility in light of counterevidence, and the nature of intertheoretic change). Notably, unlike Gallagher and apparently Zahavi, they do not make consciousness a criterion for theory possession, use, formation, or change. With such a specific conception of theories, Gopnik and Meltzoff rightly admit that not all of our knowledge is theoretical, and identify other types of cognitive mechanisms they call "modules" and "empirical generalizations." They characterize modules as possessing the same structural and functional features as theories, but unlike theories, as being resistant to counterevidence. Empirical generalizations, which include scripts, schemas, and narratives, are knowledge structures more tied to immediate experience. Because of this, they lack the abstractness and coherence of theories, and thus differ in their explanatory and predictive capacities. Gopnik and Meltzoff loosely follow Fodor (1983) on how these mechanisms are interrelated, with modules serving as input systems to "central" cognition, the place of theories and empirical generalizations. Let me be clear that I am not advocating Gopnik and Meltzoff's taxonomy of cognitive mechanisms, or the connections they draw between children's knowledge and scientific theorizing. I mention their account because it is illustrative of how theory theorists could meet the promiscuity objection without requiring theorizing to be a personal-level, conscious, reflective process. Subpersonal brain mechanisms might be

characterized as possessing representations, developing theories, etc., without their requiring personal-level forms of consciousness and intelligence.

Thus, if Zahavi wants to reject a subpersonal conception of theorizing, so social perception cannot be accounted for by TT, he must be more specific about what features of personal-level theorizing are objectionable at the subpersonal level. For example, does Zahavi understand "theorizing" as making inferences with propositionally structured representations, as in the deductive-nomological model of explanation once so dominant in the philosophy of science? Or is theorizing intended more broadly, to cover the connectionist-style processing Paul Churchland (1989, ch. 9) believes to characterize the brain's cognitive operations? Are both of these problematic forms of "theorizing"? Zahavi (2005) refers to subpersonal "*information-processing* mechanisms" (p. 222, italics added), and thus might be open to representationalist subpersonal accounts of social perception. But nothing further is said about the nature of such subpersonal representations, so perhaps even this interpretation of Zahavi's position is too quick. A nonrepresentational alternative might be found in the account of emotion perception recently argued for by Dan Hutto, and endorsed by Gallagher (in Menary 2006; see also Hutto, 2008a). Note, however, that in making this argument, Hutto directly addresses criteria for ascribing informational or representational content to subpersonal mechanisms. This is a very different kind of argument from Zahavi's phenomenology-based criticism of the pervasiveness of conscious theorizing, which is pitched at the personal level. Zahavi would similarly need to argue in a way appropriate to the subpersonal level in order to rule out subpersonal theorizing as underlying direct social perception.

As this cursory discussion shows, there are a number of interesting issues for philosophers and cognitive scientists about the subpersonal processes underlying social perception. While phenomenological evidence is important in providing adequate personal-level accounts, it alone is insufficient to rule out subpersonal theorizing as enabling social perception. As discussed in chapter 3, different kinds of evidence are needed to evaluate claims about subpersonal processes. In this case, an extended discussion of possible subpersonal explanations is required if Zahavi is to adequately evaluate TT as an account of social perception. One of the more interesting advances by theory theorists is the characterization of theories in terms of Bayesian networks (see, e.g., Tenenbaum, Griffiths, & Kemp, 2006). This has enabled the creation of computational models of cognitive phenomena such as action understanding (Baker, Tenenbaum, & Saxe, 2006). While these models remain rather simple at this point, they more precisely characterize the nature of "theorizing" in particular domains, and permit more fine-grained comparisons with human data. Accordingly, they also provide critics such as Zahavi with clearer targets against which to launch objections. While I am not here endorsing TT as an adequate characterization of the subpersonal processes enabling direct social perception (or, indeed, an adequate description of *any* subpersonal processes), my point is that Zahavi's arguments have failed to rule TT out.

3. Gallagher Against Simulation Theory

A similar dialectic about personal- and subpersonal-level processes occurs in Shaun Gallagher's (2001, 2005, 2007, 2008b) recent criticisms of ST. I will focus on Gallagher (2007), where these ideas are developed in depth for the first time. Gallagher objects to conscious simulation as an account of our everyday social experience by appeal to phenomenology. His "*simple* phenomenological argument" (2007, p. 356) is that we just do not very often find ourselves consciously simulating others' mental states. This parallels Zahavi's denial that conscious theorizing pervades our everyday social interactions. Gallagher does not deny that we sometimes engage in conscious simulation, but claims that this is relatively rare and thus cannot account for how we understand all the people we come across in our daily lives. Gallagher (2007) also objects to the claim that conscious awareness of simulation diminishes as it becomes habitual, analogous to the way our driving habits recede out of conscious awareness as we become expert drivers. This is because, Gallagher claims, even habitual processes can become objects of conscious reflection. And we do not seem capable of turning cases of social perception into ones of explicit simulation through reflection. Social perception should thus be seen as a phenomenon distinct from conscious simulation.

These objections to ST seem right as personal-level claims. Conscious simulation is phenomenologically distinct from social perception, which should be added to the list of personal-level phenomena relevant to human social understanding. This also provides an important corrective to claims about the pervasiveness of conscious simulation. If it is to account for more than the relatively rare cases of conscious imaginative simulation, ST must be treated as a theory of the subpersonal

processes underlying direct social perception. As Gallagher (2007) recognizes, phenomenological evidence will not speak to the nature of these subpersonal processes. To reject this version of ST, Gallagher thus makes a conceptual argument, contending that subpersonal processes do not meet ST's own criteria for something to be a simulation.

Gallagher (2007) begins with two definitions of "simulation" offered by the Oxford English Dictionary: (a) the *pretense definition*: "Simulation is an imitation, in the sense of something *not real*—counterfeit; to simulate means to feign, to pretend"; and (b) the *instrumental definition*: "a simulator: a model (a thing) that we can use or do things with so we can understand the real thing" (p. 359). Quoting descriptions of ST by various authors, Gallagher argues that ST combines these two definitions, such that simulation is a process where I use (i.e., "control in an instrumental way") my own psychological mechanisms as pretend or "as if" models of another person (p. 360). Gallagher then rejects the notion of subpersonal simulation by arguing that neither the instrumental condition nor the pretense condition is present at the subpersonal level.

First, with regard to the instrumental condition, if ST requires that "I (or my brain) uses or controls" a simulation, Gallagher denies that such control occurs at the subpersonal level (p. 360). The core of Gallagher's objection is that the instrumental or control condition is best understood at the personal level. If simulation is a reflective, personal-level process, it is at least partly under conscious control—something we can initiate and terminate at will. But at the personal level we do not control our subpersonal processes in this way, and thus do not "use" implicit

simulations to model the other person's mental states. Going fully subpersonal, Gallagher rejects the claim that the *brain* uses any neural processes as simulations. The objection here appears to be largely conceptual, claiming that the personal-level concepts of "use" and "control" do not make sense at the subpersonal level. Rather than simulation's being a controlled process following perception, Gallagher argues that the neural processes activated when we perceive a person are "elicited" in us. Hence, mirror neuron activation or other kinds of neural processes should not be described as simulations, but rather as part of the temporally extended process of social perception.

I'm on board with Gallagher that subpersonal processes as such are not subject to personal-level control. But it seems overly restrictive to say that one neural process cannot be "used" by another. Of course one commits the homuncular fallacy to literally apply the personal-level concept of "use" to brain mechanisms. Brain mechanisms do not use other brain mechanisms as representations or models in the same way persons understand and use external representations and models (e.g., maps, texts, scale physical models, etc.). But to imply that this is the only acceptable sense of "use" or "control" begs the question against a subpersonal version of ST, and the idea of subpersonal representation in general.

A defense of the instrumental condition for ST can be found in William Ramsey's (2007) recent critical evaluation of the appeal to representations in cognitive science. Ramsey defends precisely the subpersonal-level notion of using models as representations rejected by Gallagher. Like simulation theorists, Ramsey argues that such models serve as representations—i.e., are able to "stand in" for the things in the

world—by being structurally similar or isomorphic to the things they represent.[16] My focus here, though, is on Ramsey's discussion of how subpersonal mechanisms can be "used" by other subpersonal mechanisms without requiring those mechanisms to be intelligent homunculi. Ramsey (2007, pp. 194–203) makes his case using the example of a car navigating its way along an S-shaped track. If a real person were driving the car, one way they could steer the car would be to use a map of the track. This is an obvious case of using a model. Ramsey then considers removing the driver, turning the car into a mindless system. Could the internal workings of the car be automated such that they could still be characterized in terms of using a model of the track?

> One way we might do this, suggested by Cummins (1996), would be to convert the S-curve of the map into an S-shaped groove into which a rudder would fit. The rudder could then move along the groove as the vehicle moves forward, and the direction of the steering wheel and, thus, the vehicle's front wheels could be made to correspond to the direction of the rudder…. As the rudder moves along the groove, its change in orientation would bring about a change in the orientation of the front wheels. Because the shape of the groove is isomorphic with the curve itself, the wheels change along with the S-curve and the vehicle moves through it without ever bumping into a wall. (Ramsey, 2007, p. 198)

The car's internal workings could be characterized without any appeal to representations, as can be done (in principle) with any representational system. But the most natural way of explaining how the car navigates the road is that the groove serves as a map of the course of the track, with sections of the groove "standing in for" segments of the track. The sense in which the groove is being "used" as a model by

[16] For more on the role of models in cognition, and the sense in which models are isomorphic to what they represent, see Waskan (2006).

the other parts of the car does not require any intelligence on their part. As Ramsey summarizes, "A mindless system can still take advantage of the structural isomorphism between internal structures and the world, and in so doing, employ elements of those internal structures as representations-qua-stand-ins" (p. 200).

In the same way, brain mechanisms could "use" other brain mechanisms as models without requiring the intelligence of a person. That the task of social perception is one of understanding rather than of pure behavioral navigation should not be a barrier to characterizing the subpersonal processes in terms of simulation, of using a model. In addition, whether such a process is initiated endogenously or activated in response to stimulation from the environment does not seem relevant to whether such a process should be described as "using a model." This cuts against Gallagher's (2007) characterization of the neural processes underlying social perception not as simulations but as "effects" which are "elicited" in us by the other person's presence (pp. 360–361).

Gallagher seems to recognize this possibility of subpersonal processes meeting the instrumental condition when discussing accounts of social perception based on the motor control literature (e.g., Hurley, 2005, 2006, 2008), which apply just such an understanding of neural processes "using" other neural processes as models.[17] Gallagher does not mention any problem with the instrumental condition for these accounts, and instead argues that they fail ST's pretense condition. The core idea of

[17] Grush (2004) explicitly argues for such an account of neural representation, applying it to motor control, imagery, and perception.

these accounts is that when motor commands are sent to control the body, efference copies of these motor commands are sent to a mechanism called a "forward model" to predict the success of these movements toward achieving one's goals. One useful feature of these predictions is that they can drive corrections to behavior faster than can be done by sensory feedback. This picture of motor control is then used to offer a simulation-based account of action perception: the forward model(s) used in motor control to predict one's own movements are co-opted in social perception to simulate the actions of others and allow predictions of their behavior, and through additional processes, the mental states causing these actions. As Gallagher summarizes, these motor-based accounts claim that "the perception of the other's action is automatically informed by a sub-personal simulation; perception of action involves a loop through the [forward model]" (2007, p. 362, n. 10).

Given its uncontroversial sense of "using a model," Gallagher objects that the motor representations in these accounts fail to meet ST's pretense condition. Expressing this objection, Gallagher writes that "A specification in my motor system that the action belongs to another is not equivalent to the specification '*as if* I were carrying out the action'" and that implicit simulation requires a "representation of my own motor action *as if it were* the other's" (2007, p. 362, n. 10). As I read him, Gallagher believes that the pretense or "as if it were I"[18] component must be in the

[18] This phrasing is closer to Robert Gordan's version of ST than Alvin Goldman's. Whatever Gallagher's stance on their competing characterizations of simulation, I read Gallagher's objection to subpersonal pretense as cutting against both Gordon and Goldman on subpersonal simulation—see his discussions of Goldman (Gallagher 2007, 361) and of Gordon (Gallagher 2007, 361, n. 10). Below I

content of a representation for it to count as a subpersonal simulation. It is not enough if these states are explicitly marked as belonging to another person rather than oneself. For Gallagher, to count as simulations, they must be represented as states of oneself "pretending" to be those of another person.

As Gallagher notes, this requirement rules out treating as simulations all so called "shared representations": subpersonal representations which are agent-neutral, in that they represent a property of agents (e.g., their intentions) without specifying whether the agent is oneself or another person. Mirror neurons have been characterized as examples of shared representations since they fire both during one's own action and when observing another person's actions. If shared representations are agent-neutral, they cannot, Gallagher argues, include in their content that these are my states pretending to be another's states. Gallagher (2007) summarizes this point with regard to mirror neurons in the following passage:

> …the mirror system is neutral with respect to the agent; there is no first- or third-person specification involved. In that case, they do not register *my* intentions as pretending to be *your* intentions; there is no "as if"—there is no neuronal subjunctive—because there is no "I" or "you" represented. (p. 361)

The "motor simulations" of motor-based accounts do not have content of the kind Gallagher requires. These motor representations activated during social perception are simply the same ones activated during motor control. As normally characterized, they do not explicitly mark the agent in question (self or perceived other), so they certainly

---

attempt to undercut Gallagher's objection by appealing to Goldman's (2006) definition of simulation. I do not here delve into the differences between Gordon and Goldman on simulation, or whether Gordon would endorse such a response to Gallagher. I want to thank Marc Slors for his comments on this point.

cannot have as part of their content that this is the motor representation I would have if I were the other person. In this way Gallagher argues that motor-based accounts of social perception fail to involve subpersonal simulation. While these motor representations serve as Gallagher's primary example, he intends his point to be general: to count as a simulation, a representation must include in its content that the state in question is one of the self being used as an "as if" model of the other.

It should be noted that Gallagher's point here is that the *content* of neural states fails to meet the pretense condition. He also contends that neural mechanisms considered as *vehicles* of representational content fail the pretense condition, writing that "neurons either fire or they do not fire. They do not pretend to fire" (2007, p. 361). As Gallagher seems to recognize, ST is concerned with matters of representational content more than representational vehicles. Thus I will focus on Gallagher's claim that the content of subpersonal, neural mechanisms cannot meet the pretense condition for ST.

The pretense condition clearly applies to personal-level simulation where I consciously imagine being another person. These conscious episodes of imagination indeed include in their content that I am pretending to be the other person. But we might question Gallagher's reading of ST's pretense condition that it requires any state/process counting as a simulation to include an "as if it were I" component in its representational content.[19] Gallagher (2007) defends this reading by writing:

[19] My argument here should not be read as endorsing the standard, simulationist interpretation of motor-based accounts of social perception. There are several different versions of these models, and how they

> For [ST], a simulation is not simply a model that we use to understand the other person—theoretical models would suffice if this were all that was required. Even the fact that the model is constituted in our own mechanisms is not sufficient. Rather, I must use the model "as if" I were in the other person's situation. (p. 360)

Of course simulations must be distinguished from theoretical models. This can be done by distinguishing the kinds of representations involved. Folk psychological theorizing requires representations of folk psychological generalizations or laws: how mental states relate to environmental stimuli, behavior, and each other. It is precisely such representations which ST denies.[20] ST instead posits mental states/processes which *replicate or resemble* those of the target system being simulated. By co-opting one's own psychological mechanisms, one can represent the target's mental states without requiring the descriptive representations posited by TT. These replicated states are often characterized as "pretend" mental states, but this is not necessary: the essential feature of simulation is that the simulation replicates or is similar to the target state being simulated (Goldman, 2006). Of course this simulation process concerning another person must be distinguished from a genuine process concerning myself, and used to create explicit representations of the mental state or states as belonging to the other person.[21] As mentioned above, this is a problem for all so-called "shared

---

should be interpreted relative to TT and ST is not so straightforward. I will be developing these ideas in future work based on Herschbach (2008a, 2008b).

[20] Hybrid theory–simulation accounts, such as Goldman's (2006), do, however, acknowledge a role for representations of folk psychological generalizations. For example, "theory-driven" simulations use theoretical knowledge about the target system to generate appropriate inputs for the simulation process.

[21] How exactly this resulting state should be characterized is a matter of dispute. Simulation theorists (as well as theory theorists) usually describe this as the production of a *belief* about the other's mental state (e.g., Goldman, 2006). But phenomenologists (e.g., Gallagher, 2008b) contrast the "non-

representations," i.e., agent-neutral representations of the properties of self or other. But all simulated states need not be represented as being "as if" they were another person's. Aspects of a simulation process may go on without the identity of the agent (self vs. other) being explicitly represented. Just because the end result of this process must be categorized and attributed to an agent does not mean all aspects of the purported simulation process must explicitly represent the agent in question to count as simulations. To adapt some of John Perry's (1993, ch. 10) terminology, they may *concern* the simulated person without being *about* (i.e., explicitly representing) that person during the simulation process.

Imagine that perception of another person's, say, anger occurs by activating (a) motor representations associated with the facial expression exhibited by that person and (b) other neural states associated with the experience of anger (as Goldman, 2006, suggests is actually the case). Following many characterizations of ST, it makes sense to say that the "online" function of these states is for my own experiences of anger, and that they are being used "offline" for understanding another person's anger.[22] Because of this, these "shared representations" would arguably replicate or resemble those of the perceived angry person. While these states do not *represent* "the emotional state I would be in if I were that other person," they are the states *I would be*

---

conceptual" experience of direct social perception with the "conceptual," belief-based understanding of reflective simulation and theorizing. Although this issue seems important to adequately describing the personal-level phenomena of social understanding, it requires delving into the thorny issue of the nature of concepts, and is thus beyond the scope of this chapter.

[22] Note that the "online–offline" distinction used in defining simulation is not the same as the one I am using in this dissertation to distinguish types of personal-level phenomena.

in if I were myself angry. While such states must still be used to create a distinct representation *that he is angry*, the fact that I would be in these states if I were myself angry is all the "pretense" needed by ST. Against Gallagher (2007, p. 361), I cannot see how creating this extra representation (i.e., an explicit mental state attribution) requires the stronger, personal-level sense of pretense. Further, this weaker sense of "as if it were I" is sufficient to distinguish it from TT accounts. As Goldman (2006) describes, TT would most likely posit that a mechanism *separate* from the ones involved in my own emotional states is responsible for my attributions of emotions to others (p. 114). And to count as theoretical, such a mechanism would likely represent three kinds of information: visual information about bodily states and movements, particularly of the face; information about how these bodily states correlate with different types of emotional states; and information about the typical environmental elicitors and behavioral effects of particular emotions (Goldman, 2006, p. 119). ST is negatively defined as denying such information-rich mechanisms; a more positive definition focuses on the reuse of our own mechanisms to simulate those states of the target agent (Goldman, 2006, p. 34). This is just what the case of emotion perception described above involves, since the person's own emotion mechanisms are co-opted for understanding the other's anger.

Overall, I believe Gallagher loses some important explanatory purchase by confining "simulation" to the personal level. Admittedly, ST developed as a description of conscious simulation, and the personal and subpersonal levels have been inadequately distinguished in many discussions of ST. But denying simulation a subpersonal role seems to obscure the role of states/processes which

replicate/resemble those of the perceived target, which I would be in "if it were I" in the position of the target. And it obscures how such subpersonal processes differ from those involving "theoretical" representations. This contrast seems to be what much of the literature distinguishing TT and ST has concerned itself with. The issue is not about the use of the term "simulation." What is important is that the theoretical distinctions made in the literature on ST seem to have import at the subpersonal level, and should not be lost so as to avoid confusions between personal and subpersonal levels.

Additionally, there remains the question of how personal-level, conscious simulation must be characterized at the subpersonal level. It is an open possibility that conscious, imaginative simulation and direct social perception, which in phenomenological terms are distinct personal-level phenomena, are driven by the same or similar subpersonal mechanisms—ones which may be understood in terms of subpersonal simulation. If this were the case, simulation would play a much greater role in our social understanding than Gallagher suggests. Further, it would raise questions about the phenomenology-based distinction between these personal-level phenomena, and more generally about the presence or absence of consciousness in characterizing cognitive phenomena. Phenomenologists want to draw a firm distinction between conscious, imaginative simulation and direct social perception on the basis of phenomenology. Being enabled by the same subpersonal mechanisms might lead us to seek better phenomenological descriptions of these personal-level phenomena, emphasizing similarities between conscious simulation and direct social perception rather than simply their differences. More generally, this case would raise hard

questions about how we categorize phenomena at different levels, and whether any particular level should be privileged. For example, I have been assuming that cognitive phenomena should be identified at the personal level, then explained in terms of subpersonal-level mechanisms. Subpersonal-level mechanisms are then defined in terms of the personal-level phenomena that they enable. But should explanatory priority be given to the personal level, in particular to features of conscious experience? One might instead treat the subpersonal-level processes as fundamental, and categorize cognitive phenomena at this level. The presence or absence of consciousness would then be relatively unimportant to typing cognitive phenomena. These are thorny theoretical issues for which I do not have any ready answers. I mention them as important conceptual issues left for future research into the phenomenology and subpersonal mechanisms of social understanding.

4. Conclusion

What we are left with is that these phenomenological critics have not provided compelling objections to TT and ST as accounts of social perception. We certainly need to respect the phenomenology, by acknowledging a distinction between (a) directly perceiving people's mental states, and (b) attributing mental states via reflective processes of theorizing and simulation. There is certainly descriptive work left to be done on this front. For example, how are the products of direct perception versus reflective cognition related? Is the direct perception that someone is, say, angry the same as coming to that conclusion via theorizing or simulation? Or is the direct

perception of anger different in some way?[23] There is also the issue of distinguishing the various mental state types. Emotions and intentions are most often identified as being directly perceived. But what about states less connected to behavior, like belief? We certainly seem to be able to non-inferentially attribute beliefs to people. How is this different from the direct perception of emotions or intentions?

These are questions befitting phenomenological investigation. But even once we obtain such answers, there remains much about social perception left to account for: namely, the nature of the subpersonal processes enabling this perceptual experience of humans as embodied agents with sensations, emotions, intentions, etc. I have argued that Zahavi and Gallagher's reasons offered against characterizing these subpersonal processes in terms of theorizing or simulation are unconvincing. Zahavi's arguments are inadequate mainly because he does not directly address what is problematic about the notion of subpersonal theorizing. While Gallagher directly argues against the notion of subpersonal simulation, I have attempted to show why these arguments are unpersuasive.

Perhaps the thrust of Zahavi and Gallagher's criticisms is correct, that we need conceptual development to properly characterize the subpersonal processes underlying social perception. This may very well be the case (see Bechtel, 1994, 2005). In this chapter I have remained uncommitted about whether TT and ST are appropriate subpersonal accounts of social perception, or whether alternative descriptions are

---

[23] See footnote 21.

needed. But to rule out TT and ST as contenders, what is required is a more detailed discussion of appropriate subpersonal explanations than what has been offered thus far by Zahavi and Gallagher.

Chapter 4 is a reprint, with slight modifications, of "Folk psychological and phenomenological accounts of social perception," *Philosophical Explorations*, *11*(3), 223–235. Permission to reproduce this material has been granted by the copyright owner, Taylor & Francis. The dissertation author was the primary investigator and author of this paper.

Chapter 5. Defending the Pervasiveness of Belief–Desire Psychology

1. Introduction

The second challenge to the folk psychological picture of human social understanding I described in chapter 1 questions the scope or pervasiveness of folk psychology in human social understanding.[24] The phenomenological critics challenge the idea that mental state attributions via theorizing or simulation procedures "constitute the primary way in which we relate to, interact with or understand others" (Gallagher, 2001, p. 85). While admitting that we sometimes consciously, explicitly reflect on people's mental states, they claim that "such instances are rare…relative to the majority of our interactions" (Gallagher, 2001, p. 85), which they contend do *not* involve mental state attribution. In the terminology I prefer, this second challenge of the phenomenological critics can be summarized as the claim that *online social understanding is not driven by folk psychological reasoning*. In other words, the phenomenological critics question whether we need to attribute mental states at all in order to unreflectively interact with other people. Granted the assumption that unreflective interaction occurs at least as often as conscious reflection (the phenomenological critics seem to assume it occurs much more often), this would mean folk psychology is much less pervasive in our daily lives than is assumed by the

[24] Gallagher (2001, 2005) calls claims about the scope of folk psychology "pragmatic claims," as opposed to "developmental claims" about the development of our folk psychological abilities.

traditional folk psychological accounts, which treat mental state attribution as driving all or at least most of human social understanding—whether our conscious reflections about other people or spontaneous, unreflective interactions with them. Since the phenomenological critics tell a distinct story about the attribution of emotions and intentions (as we saw in chapter 4, they think we can directly perceive these states), this second challenge is best seen as specifically targeting the role of *belief and desire attribution* in online social understanding.[25] As described in chapter 1, this is a standard assumption of the folk psychological picture: Mature folk psychology is often characterized as "belief–desire psychology," and is usually ascribed to a person when they can reliably pass false-belief tasks, the standard measure for the mature concept of BELIEF.

As with their discussions of social perception, phenomenological description of our first-person experience serves a fundamental role in this second challenge. The critics note that much of our everyday social understanding does not involve conscious belief–desire attribution. I accept this personal-level description of our conscious experience. The question then is whether belief–desire attribution plays a role in the *subpersonal-level processes* enabling online social understanding. In chapter 4 I argued for the possibility of subpersonal versions of TT and ST, which allows for the possibility of mental state attribution occurring outside of our conscious awareness. This opens up space for folk psychology to characterize online social understanding.

[25] Gallagher (2001, 2005) is explicit in treating "mind-reading" as involving the attribution of beliefs and desires. Ratcliffe (2006a, 2007, 2008) similarly isolates belief–desire attribution as the central aspect of the folk psychological picture which he rejects.

But is there any evidence to actually support this? Or is it simply an assumption about the pervasiveness of folk psychology, reading folk psychology into behaviors which could be explained by simpler subpersonal processes?

To address this issue, I will focus one central aspect of our mature folk psychological abilities: false-belief understanding. False-belief understanding is widely recognized as a paradigmatic case of *mentalistic* understanding because it requires appreciating the *representational* nature of beliefs. If it can be shown that people display an understanding of others' false beliefs in online contexts, rather than in cases of reflective, offline cognition, it would serve as strong evidence against the phenomenological critics' skepticism about the role of folk psychology in unreflective, online social interaction. In this chapter I will provide empirical evidence that false-belief understanding can indeed be expressed in our online responses, thus arguing that the phenomenological critics have given too limited a role to mental state understanding in our everyday, unreflective, online social understanding.

In section 2 I will describe more clearly the phenomena of interest to the phenomenological critics, which I will call cases of "online" social understanding. I will fill out what I mean by the distinction between "online" and "offline" social understanding, and why I prefer this terminology to the phenomenological critics' own descriptions of these phenomena. I will also show that the phenomenologists are not the only ones emphasizing these phenomena. Empirical researchers of various stripes have recently come to recognize the methodological limitations of previous research on social understanding—specifically, that it overemphasized offline social understanding and failed to adequately address online forms of social understanding.

In section 3 I will describe an alternative to belief–desire psychology emphasized by the phenomenological critics, which can be called *situational understanding*, as well as further work in this area by researchers outside the phenomenological tradition. While these alternatives to folk psychology should be included in an enriched picture of human social understanding, I will argue that their mere existence and the arguments offered in favor of them do not establish the phenomenological critics' more radical claim that folk psychology plays no role in online social understanding. In sections 4–5 I will provide reason to think folk psychology does play such a role. In section 4 I explain why false-belief tasks are widely recognized as tests of genuine folk psychological understanding, and fill out the threat that online false-belief understanding poses to the phenomenological critics. In section 5 I will describe evidence from experimental research that we are indeed able to make false-belief attributions in the context of online social interactions. In section 6, I will defend my interpretation of this evidence against possible objections from the phenomenological critics.

2. Online Versus Offline Social Understanding

As I described in chapter 1, the phenomenological critics often introduce this second aspect of their challenge to folk psychology in the context of criticizing "theory of mind" experiments like the standard false-belief task (Gallagher, 2001, 2005; Ratcliffe, 2007, ch. 4). Such tasks require a child to *observe* someone's behavior, and *predict* or *explain* their behavior by attributing propositional attitudes to

them. The phenomenological critics point out that "theory of mind" tasks such as this one place the child in the role of *theorist*, providing explanations and predictions of a third-party's behavior. Even if the child is not required to provide verbal explanations and predictions, the child is at least required to somehow report to the experimenter a behavioral prediction, perhaps by pointing to a location. It is precisely this explicit explanation and prediction of behavior based on mental state attribution which the two dominant folk psychological accounts, TT and ST, were developed to explain. But such episodes of reflective prediction and explanation fail to capture the full range of ways in which we relate to other people—specifically, our *unreflective interactions* with other people are left out. For example, even in "theory of mind" tasks the child must be able to interact with the experimenter in various ways. Yet the folk psychological accounts focus on the ability to reflect on a third-party's behavior, rather to participate in social interactions. It is at this point that the phenomenological critics question the applicability of TT and ST to such participatory phenomena: If our evidence for TT and ST come from "theory of mind" experiments, how are we to know if mental state understanding is even required for unreflective social interaction?

Before I respond to this criticism, I want to get a better picture of the phenomena the phenomenologists think (a) are left out of the folk psychological picture, and (b) need not be explained in terms of TT and ST, or the use of mental state attribution at all. The distinction between theoretical and participatory social understanding to which the phenomenological critics want to call attention can be helpfully understood in terms of Wheeler's (2005) distinction between "online" and "offline" intelligence, which he introduces while developing a Heideggerian

conceptual framework for cognitive science. Essential for my purposes is that Wheeler explicitly pitches this as a distinction among phenomena at the *personal level*.[26] Online intelligence involves an organism's active sensorimotor engagement with the world: "A creature displays online intelligence just when it produces a suite of fluid and flexible real-time adaptive responses to incoming sensory stimuli" (p. 12). Offline intelligence, in contrast, is exhibited when an organism is not acting, but rather reflecting on the world and its possible actions. This is not primarily a contrast between psychological processes that are explicit and available to consciousness, and ones that are not. Rather, it is about the stance a whole organism takes toward its environment—online sensorimotor interaction versus disengaged contemplation—which makes it especially clear why this is a distinction at the personal level rather than the subpersonal level.

But just focusing on the bodily behavior of a person is not sufficient to capture the distinction phenomenologists want to draw here. We must look to the other class of personal-level phenomena I've discussed: conscious experiences. The paradigm cases of online and offline understanding clearly do differ this regard. Online

[26] Wheeler (2005), however, talks of "agential" and "subagential" levels rather than "personal" and "subpersonal" levels, to emphasize that the distinction "is applicable to any creature that competently inhabits its environment—person, human, or otherwise," and because he thinks "the language of agents…carries less philosophical baggage" (p. 300). Wheeler also draws on McDowell (1994), as I did in chapter 2, when characterizing his Heideggerian account of the relations between levels. But there are differences between Wheeler's account and my own. A central one is that Wheeler enthusiastically endorses vehicle externalism as an account of the subpersonal-level states and processes which might enable some personal-level phenomena—in particular, cases of online intelligence. I am more skeptical about the need for vehicle externalism, as I discussed briefly in chapter 2. So for my purposes here, I am only adopting Wheeler's delineation of personal-level phenomena in terms of online and offline intelligence, without endorsing Wheeler's views on how these phenomena might be explained at subpersonal levels.

intelligence, such as skillfully hammering a board, involves little conscious awareness: instead of consciously experiencing all of the features of the environment to which you are responding and consciously thinking about how to respond to them, you simply "cope" with the situation using your implicit know-how. On the other hand, offline reflection more paradigmatically involves conscious experience. To continue the hammering theme, you might consciously imagine possible things to build with the materials you have on hand, and subvocally talk through the pros and cons of each. Of course we at times reach such decisions without much conscious reasoning, if any at all (we're all familiar with cases where our offline intelligence is simply a matter of pausing long enough for the solution to simply pop into our consciousness). And surely even paradigmatic episodes of conscious, offline reflection also require many cognitive processes occurring outside of conscious awareness. So the distinction between online and offline intelligence clearly cannot capture all the features relevant to characterizing cognition. But this contrast between online and offline stances is useful for capturing the distinction in forms of social understanding made by the phenomenological critics.

While standard "theory of mind" tasks require the child to provide a prediction or explanation to the experimenter, and thus involves interaction, traditional folk psychological accounts clearly focus more on *offline* forms of social understanding, where we are *thinking* about other people's behavior and making explicit *judgments* about their mental states—in the case of standard false-belief tasks, using mental state concepts to think about someone's false belief to explain or predict their behavior.

What the phenomenological critics want to call our attention to are *online* forms of social understanding, such as the child's active engagement with the experimenter.

This distinction between online and offline social understanding is sometimes characterized by phenomenologists as the distinction between "second-person" and "third-person" understanding—understanding someone as a "you" as opposed to as "he" or "she"—and sometimes as between "theoretical" and "pragmatic" stances (Ratcliffe, 2007, ch. 6). These are often combined, contrasting "third-person theorizing/observation" with "second-person interactions" (Gallagher, 2005). I prefer Wheeler's terminology because I believe it better carves up the relevant phenomena. The online–offline distinction captures the majority of cases the critics want to contrast, while showing what is in common between cases which would be treated as distinct using the alternative vocabulary. For example, I can engage in "offline" reflection on the behavior of, and can act "online" in response to, both second- and third-persons; while sometimes it is relevant to mark the distinction between activity in response to second- vs. third-persons, the critics more often intend to contrast online interaction with offline reflection, than to distinguish whether the person I am reflecting on is a "you" or a "he"/"she." Furthermore, the online–offline distinction remains neutral with regard to the kinds of psychological processes underlying such activity—e.g., sensorimotor vs. "theoretical," conceptual processes. To call all reflection "theoretical" in nature just begs the question against ST unnecessarily. And to contrast sensorimotor behavior with "theoretical" cognition is to miss the ways in which action can be mediated by theoretical knowledge.

Gallagher (2005) describes the kinds of online phenomena the phenomenologists have in mind as "embodied practices—practices that are emotional, sensory-motor, perceptual, and non-conceptual" (p. 224). Gallagher's language here is representative of the phenomenological critics' proposal that these capacities for online social interaction are not amendable to description in the languages of TT or ST, which were developed to explain offline forms of social understanding. As examples of these "embodied practices" Gallagher lists "imitation, intentionality detection, eye-tracking, the perception of meaning and emotion in movement and posture, and the understanding of intentional or goal-directed movements in pragmatic contexts" (p. 230). It is not especially clear how all of these are specific to online social *interaction*, as De Jaegher (2006, 2007) has noted. It is for this reason that I analyzed the phenomenologists' account of social perception—which could in principle be used in the course of interacting with "second-persons" or passively observing "third-persons"—separately from their claims about online social interaction.

Clearer cases of online social interaction come in Gallagher's (2005) discussion of what Trevarthen and Hubley (1978) call "secondary intersubjectivity." "Secondary intersubjectivity" is a name for the "triadic" person–person–object interactions which human children begin to engage in when they are around 9-months of age. Gallagher (2005, p. 228) quotes Peter Hobson's summary of this notion:

> The defining feature of secondary intersubjectivity is that an object or event can become a focus *between* people. Objects and events can be communicated about. …the infant's interactions with another person begin to have reference to the things that surround them. (Hobson, 2002, p. 62)

Secondary intersubjectivity is well exemplified by the cooperative games children engage in with adults starting around age 2. Warneken, Chen and Tomasello's (2006) experiments, for instance, involved various social games where two players perform complementary roles. One of their tasks, called the "double tube" task, involved an apparatus with two tubes arranged horizontally on a decline, such that an object placed into a tube at the higher end would slide down and come out the opening at the other end. The tube was long enough such that the same person could not both place an object in the opening of the tube and catch it on its way out the other end. The experimental task thus required a child and an experimenter to cooperate to slide objects down the tube: one person would send a wooden block down one of the tubes from the higher end, and the other person at the bottom end of the tube would catch it in a tin can, which made a rattling sound. Obviously this is a very simple game, but one the 18- to 24-month-old children in this study found interesting enough (in contrast to the young chimpanzees they also studied, who never seemed motivated to cooperate in such a game). These children often picked up the nature of the game from a single demonstration by the experimenters, and readily interpreted the experimenter's nonverbal invitation to partake in the game (namely, when the experimenter took up one of the roles, say, holding the can at the bottom of one the tubes, and alternated gaze between the child and the block the child would need to pick up and place inside the tube). Without taking a stand on exactly what each person understands about the other, playing such games is unquestionably a dynamic social interaction involving the interpretation and coordination of various bodily behaviors

and nonlinguistic auditory expressions. This is very different from the situation of children in the standard "theory of mind" tasks described above and in chapter 1, where instead of actively interacting with another person, the child is told a story about some third party and must explain or predict that person's behavior.

As suggested by this experimental study, phenomenologists are not alone in recognizing the disconnect between online social phenomena and the standard experimental paradigms using offline tasks. For example, Slaughter and Repacholi (2003), in their recent introduction to a volume on individual differences in "theory of mind" (i.e., folk psychology), comment on just this issue:

> Given the need to assess theory of mind more comprehensively, it seems worthwhile to ask the question: *What is it that we do in our everyday social reasoning that is different from what we assess with standard and higher level mental state attribution tasks?* Several dimensions of difference between laboratory theory of mind tasks and everyday social reasoning spring to mind. For instance, in everyday mind reading, we compute mental states *online*, and often *act* on these computations. It would seem rare for us to explicitly *reflect* on the mental state attributions we make in the course of social interactions; instead, we are much more likely to act on those attributions with an *immediate behavioral or linguistic response*. (p. 7, italics added)

While the phenomenological critics might disagree with some of their language (e.g., the focus on "everyday social *reasoning*," and the assumption that mental state attributions are so essential to social interaction), Slaughter and Repacholi share their concern that there is a methodological problem with the way social understanding is studied under the heading of folk psychology or "theory of mind." Addressing both theory and methodology, Carpendale and Lewis (2006) explicitly chose the term "social understanding" rather than "folk psychology," "theory of mind," "mindreading," "mentalizing," or "belief–desire reasoning" to characterize the topic of

their book—even though, following the literature, they mainly focus on children's understanding of the mind—because at times "these terms come attached to theoretical assumptions regarding the nature and development of children's social knowledge—assumptions [Carpendale and Lewis] believe need to be examined" (p. xi). One such theoretical assumption Carpendale and Lewis identify and try to undermine is the "individualistic perspective" of traditional views; they argue in contrast that accounts of the development of social understanding must give a fundamental role to social interaction. Thus, some psychologists themselves are beginning to recognize the narrowness of research and theory into human social understanding, that social interaction has specifically been given short shrift in the folk psychological framework.

This methodological worry is also beginning to be recognized in the neuroimaging literature, where some researchers have recently begun using online rather than offline tasks of social understanding. Approximating situations where participants do not themselves act toward another person, but where the other person directs some action toward the participant, Oberman, Pineda, and Ramachandran (2007) and Schilbach et al. (2006) used video clips where the subject's passive participation in social interaction is implied. The former study used videos of people tossing a ball to one another, and one condition involved the ball being passed in the direction of the viewer, as if the viewer were a participant in the game. The latter study used videos of fairly realistic virtual humans making facial expressions directed either to an unseen person to the left or right of the viewer, or directly at the viewer. The use of recorded video in these studies means social interaction is only one-way:

the person in the video directs attention to the participant, but the participant does not provide a response. Other studies engage participants in more active social interaction—although the kind of active responses possible given current scanning technology is fairly limited. "Rock, paper, scissors" (H. L. Gallagher, Jack, Roepstorff, & Frith, 2002) and reciprocal exchange games (McCabe, Houser, Ryan, Smith, & Trouard, 2001) against human and computer opponents are examples of what has been attempted thus far. Spiers and Maguire (2006) attempted to provide more rich stimuli by using virtual reality environments, in this case a first-person driving game set in London. These studies illustrate that researchers are attempting to make the stimuli and behavioral tasks used in neuroimaging studies better approximate real-world social interactions requiring online responses.[27] This way we can be more confident in making claims about the subpersonal mechanisms enabling everyday

[27] Iacoboni et al. (2004) raise concerns about traditional, offline "theory of mind" tasks from another direction: specifically, that the traditional studies compare their experimental condition to an active control task (e.g., comparing the reading of stories highlighting people's mental states vs. stories highlighting physical causes), rather than to a true resting state. Whereas the standard methodology leads, e.g., to the claim that "theory of mind" tasks involve *increased* activation of the medial prefrontal cortex (MPFC), comparing brain activation during "theory of mind" tasks to a true resting state may show *deactivation* of the MPFC. Combining this possibility with their finding of increased MPFC activity when using realistic social stimuli (i.e., videos of everyday social interactions, as opposed to using vignettes read by participants, shown without any instructions about what they should attend to), Iacoboni et al. suggest that performance on "theory of mind" tasks may involve different neural processes than those activated when observing realistic social stimuli. While this hypothesis depends on the speculation that traditional (i.e., story-based) "theory of mind" tasks will show deactivation of the MPFC (based in the work of Mitchell, Heatherton, & Macrae, 2002), it raises important questions about the extent to which the processing of social information is "part of the brain's default state circuitry" (Iacoboni et al., 2004, p. 1171). While focusing more on the methodology used in interpreting neuroimaging results than simply the types of tasks used, this criticism falls in line with the concern of the phenomenological critics that the standard experimental tasks may not tell us much about everyday social understanding.

online social understanding, as compared to generalizing from studies using only offline tasks.

3. Beyond Belief–Desire Psychology

In highlighting the contrast between online and offline forms of social understanding, the phenomenological critics have clearly identified a relevant distinction amongst the phenomena of human social understanding to which researchers must pay attention, and increasingly are doing so. To summarize, the phenomenological critics want to: (a) deemphasize the significance placed on "theory of mind" experiments such as standard false-belief tasks, since they only focus on offline forms of social understanding; and (b) place greater importance on online social understanding. In this respect, they are in line with empirical researchers, as I've indicated above, that our personal-level descriptions of the phenomena of human social understanding must be enriched. Yet the phenomenological critics go further than this, claiming (c) that online social intelligence should not be interpreted as involving mental state attribution. It is this more radical claim that I believe cannot be justified.

The critics convincingly argue from considerations of phenomenology that folk psychological accounts have overstated the importance of conscious mental state attribution to our everyday navigation of the social world. Their work here is not, however, entirely negative. They offer alternative accounts of what psychological processes besides belief–desire reasoning might be driving our online social

understanding. In the following section I will describe one alternative discussed by the phenomenological critics, which I will call *situational understanding* (following Gallagher, 2004; Ratcliffe, 2007). I will then appeal to work by researchers outside the phenomenological tradition which says more about the nature of situational understanding, and develops a few other alternatives to belief–desire psychology.

For example, the phenomenological critics suggest that we often understand people's behavior in terms of the shared situations and social roles people inhabit, rather than in terms of their mental states (see especially Gallagher, 2004; Gallagher & Zahavi, 2008, ch. 9; Ratcliffe, 2007; Zahavi, 2005, pp. 163–168). A favorite example (e.g., Ratcliffe, 2007, ch. 4) is that we understand a waiter's actions and can interact with him not because we interpret his mental states, but because we understand his social role as a waiter, in relation to our own as a customer. Such situational understanding is a matter of understanding how people normally act in particular situations with particular social roles.

From outside the phenomenological tradition, Bermúdez (2003, 2005) and Maibom (2007) have made similar points about forms of human social understanding that do not involve belief–desire attribution. Bermúdez (2003, 2005) argues that much of our social interaction can be enabled by the use of heuristics and script- or frame-based knowledge. The heuristics Bermúdez has in mind are strategies for social interaction that only require recognizing how other agents have behaved, without any appreciation of their mental states. His core example is the "TIT-FOR-TAT" algorithm for how to act in the indefinitely repeated prisoner's dilemmas characterized by game theorists. This strategy simply says: start out cooperating, and then do

whatever your interactive partner does. Applying this heuristic does not even require any prediction or explanation of others' behavior, let along mental state attribution. It simply requires behaving in certain ways in response to how another agent has behaved. A second kind of social understanding not based in folk psychology identified by Bermúdez is precisely the understanding of routine situations and social roles described by the phenomenologists. He even uses the same example of interacting with a waiter, arguing:

> It would be too strong even to say that identifying someone as a waiter is identifying him as someone with a typical set of desires and beliefs about how best to achieve those desires. Identifying someone as a waiter is not a matter of understanding them in folk psychological terms at all. It is understanding him as a person who typically behaves in certain ways within a network of social practices that typically unfold in certain ways. The point is that this is a case in which our understanding of individuals and their behavior is parasitic on our understanding of the social practices in which their behavior takes place. (Bermúdez, 2005, pp. 203–204)

Bermúdez suggests that situational understanding should be explained (I would say, at the subpersonal level) in terms of "frame-based forms of knowledge representation," which permit similarity- and analogy-based reasoning, rather than in terms of a theory or set of rules. Rather than having "general principles about how social situations work," we "rather have a general template for particular types of situation with parameters that can be adjusted to allow for differences in detail across the members of a particular social category" (2005, p. 204). So on this account, we would have a frame-based representation for the situation of ordering food at a restaurant containing information about how the person playing the waiter role is expected to behave and how I in the customer role should act. The present situation will be matched against

the prototype represented in the frame, and any differences from the prototype—due to say, the type of restaurant (e.g., fine-dining versus more causal restaurants)—will be handled by analogical reasoning. What sets this account apart from folk psychological accounts is not the appeal to frame-based representations or similarity- and analogy-based reasoning, as opposed to the theoretical representations and reasoning of TT or the simulation processes of ST. Rather, what makes Bermúdez's discussion of situational understanding an alternative to folk psychology is that it would not "include specifications of the mental states of the other parties in the interaction" (2005, p. 205). By at least sketching an account of the kind of subpersonal-level mechanisms responsible for situational understanding, Bermúdez offers a more developed account of situational understanding than that provided by the phenomenological critics.

Maibom (2007) similarly tries to give an account of the knowledge structures and psychological processes involved in situational understanding. She differs from Bermúdez, however, by arguing that we should explain situational understanding in terms of theoretical models, rather than heuristics, scripts, or frames. Here "models" should be understood in the sense used in the philosophy of science as a replacement or supplement to theories (i.e., laws or universal generalizations). Models are abstract objects consisting of "sets of objects with relations, properties, and functions defined over them," which in order to represent real things in the world "must be supplemented by so-called theoretical hypotheses, specifying the respects in which and degree to which they fit the world" (p. 567). Maibom defines "social models" as "models of social structures and institutions and the individuals within them…

individuals occupy roles, and the way that they interact with others is a function of their role and the role of the other person(s) [in a social structure or institution]" (p. 568). She explains our familiar example of the interaction between waiter and customer in terms of possessing "restaurant models" and the ability to apply them. As she first developed in Maibom (2003), Maibom (2007) argues that folk psychological understanding should also be understood in terms of theoretical models, as opposed to the rule-based theories of traditional TT or simulation processes of ST. Because we do not need to appeal to representational mental states to talk about the purpose or function of a social institution (e.g., that it is the purpose of a school to impart knowledge) and the roles played by agents in that institution (e.g., students are there to learn and teachers are there to teach), social models are thus distinct from folk psychological models.

In addition to social models, Maibom (2007) describes another class of models important to human and nonhumans social understanding not involving the attribution of representational mental states, which she calls "behavioral models." Models of behavior characterize agents "as behaving in goal-directed ways, having goals, and standing in perception-like relationships to their environment" (p. 559). Maibom argues that we can understand behaviors as goal-directed without treating them as actions caused by representational mental states, as folk psychological models do. Rather, we can type behaviors by the motion properties of agents relative to environmental objects. Further, Maibom argues that others' goals and perceptions can be understood nonmentalistically as extensional relations between organisms and objects or features of the external environment (analogous to the account of simple

desires I will give below in section 4). In this way, models of behavior are claimed to permit skilled social interactions without the attribution of mental states.

Maibom's talk of models in the scientist's sense might imply that model-based knowledge might only apply to offline, reflective cognition. This is not, however, Maibom's intention. She claims that many of the hypotheses by which we apply these models "are best understood as implicit, even embodied knowledge ([i.e., consisting in] motor programs)" (p. 574). Maibom thus posits model-based knowledge to account for online social understanding. Further, when she says "Knowledge of hypotheses has a lived quality to it that is not explicitly represented by the organism, but is nevertheless reflected in its behavior" (p. 574), it suggests that her account of behavioral, social, and folk psychological models and the hypotheses by which we apply these models is best understood as being made at the subpersonal rather than the personal level. While Maibom does not give any details about how model-based reasoning is supposed to be implemented in a physical system like the brain, it is an interesting alternative to the theoretical reasoning of TT and simulation procedures of ST.

In sum, Bermúdez and Maibom are in line with the phenomenological critics in that they think human social understanding is not all based on folk psychology. With this I am in agreement. We need to both better delineate the personal-level phenomena of human social understanding—for instance, by noting the distinction between online and offline intelligence—as well as enrich our subpersonal-level accounts to include processes not involving folk psychology, i.e., ones not involving mental state attribution (where mental states are understood as representational states). But merely

sketching possible alternatives to folk psychology does not mean we in fact use them often. We might very well complement these nonmentalistic modes of social understanding with folk psychology. Why should we think folk psychology is less pervasive than traditionally assumed?

Beyond the *phenomenology-based argument* of the phenomenologists, Bermúdez and Maibom give us a few additional reasons to think at least some, and perhaps many, aspects of human social understanding do not involve folk psychology. I'll start with Bermúdez's (2003, 2005) *computational complexity argument* against the pervasiveness of folk psychology. He argues that alternatives to folk psychology are likely to be needed because of the computational complexity of attributing beliefs and desires to other people, particularly when trying to explain and predict the behavior of multiple interacting agents. Bermúdez thinks such cases would

> require each participant to make predictions about the likely behavior of other participants, based on an assessment of what those participants want to achieve and what they believe about their environment. For each participant, of course, the most relevant part of the environment will be the other participants. So, my prediction of what another participant will do depends upon my beliefs about what they believe the other participants will do. The other participant's beliefs about what the other participants will do are in turn dependent upon what they believe the other participants believe. And so on. (2005, p. 195)

This use of folk psychology in cases of multi-agent interaction would result, according to Bermúdez, in "a computationally intractable set of multiply embedded higher-order beliefs about beliefs" (2005, p. 196). Bermúdez argues that this computational complexity objection cuts against both TT and ST. Focusing first on TT, this objection does not claim that TT could never in principle describe the kind of theoretical reasoning that would be required for multi-agent interaction. Bermúdez identifies

game theory as providing the formal tools a proponent of TT would need to characterize the folk psychological reasoning involved in such cases. Rather, Bermúdez claims that "What thinking about computational tractability should do…is at least to cast doubt upon whether [TT] could be a correct account of the form of social understanding in the vast majority of situations" (2005, p. 195). Bermúdez raises similar doubts about ST, pointing to the computational complexity of engaging in multiple simultaneous, independent simulations of the mental states of multiple interacting agents. Thus, the heart of Bermúdez's objection is that using folk psychology (whether by means of theorizing or simulation processes) seems too computationally complex to be used for online social interaction, particularly when interacting with multiple agents.

While such doubt may be warranted, it is certainly not a decisive objection against folk psychology's playing a pervasive role in human social understanding. Bermúdez does not offer a formal proof of the computational intractability of folk psychological reasoning for multi-agent interaction. Without knowing more about the computational abilities of human brains, it is too quick to say such folk psychological reasoning is beyond their computational capacity. Further, in response to the computational complexity objection, Bermúdez seems to jump from full-blown folk psychological reasoning straight to nonmentalistic forms of social understanding. It would seem that there could be a middle ground where mental state representations are used in a way simpler than depicted by the full-blown folk psychological reasoning. Such heuristic applications of mental state representations might fail to reflect some of the complexity of the mental states of our interactive partners—e.g., I may not be able

to represent all the ways one of my interactive partners beliefs reflect the beliefs of the other agents we are interacting with—but work well enough to guide our social interactions. The folk psychological picture does not require that we are accurate in our mental state attributions, just that our online social understanding is enabled by mental state attribution of some sort or another. Thus, accepting the computational complexity argument against full-blown folk psychological reasoning does not rule out the possibility that computationally simpler forms of folk psychological reasoning play a pervasive role in human social understanding.

Another consideration offered by Maibom (2007) is that behavioral and social models might have explanatory and predictive power distinct from what folk psychology can provide. In the following passage, Maibom argues that social models alone may be able to guide some social interactions, without the use of folk psychology:

> Consider an everyday transaction like paying for gas at the gas station. When I enter to pay, folk psychological models are not particularly useful in helping me figure out how to interact with the store attendant. Imagine that after having handed him my credit card, he hands me a slip without saying anything. I know that people usually want to do what they do, so I can be pretty certain that he wants me to have the slip. I might also attribute to him the belief that by producing that motion, he is giving me the slip. Even with all this information, I am not in a particularly good position to figure out what to do next. This is *not* because I have inferred beliefs and desires that are irrelevant to the situation, but because without the requisite knowledge of credit card interactions, I cannot frame his behavior at the level of description that is useful for me to figure out what *I* should do. But whereas I will have difficulties figuring out what to do without acquiring knowledge of a relevant social model, I can get by without the application of folk psychological models. (Maibom, 2007, p. 573)

I accept Maibom's point that since behavioral, social, and folk psychological models each focus "on a different aspect of subjects and what they do—how an organism

relates to its environment, what internal events cause an action, or the role that a subject plays in a social interaction—they each provide a different understanding of the situation," and thus on their own differ in their explanatory and predictive power (p. 572). But this does not establish that humans do not pervasively use folk psychological models. Even if there are tasks of social navigation for which behavioral and social models seem sufficient, it is possible humans in fact apply their folk psychological in such cases. As Maibom suggests, our folk psychological models might very well need behavioral and social models to "constrain the space of possible beliefs and desires" I might attribute to an interactive partner (p. 573). This does not, however, establish that humans often navigate online social interactions by means of behavioral and social models alone.

What seems to be grounding Maibom's view here is a basic appeal to parsimony. If these nonmentalistic alternatives exist, why not think they work alone to drive our online social interactions, rather than being complemented by folk psychology? This approach might be supplemented by a comparison of human social understanding with that of nonhuman animals. If folk psychological reasoning is evolutionarily and developmentally late to emerge, why not assume we often use the nonmentalistic forms of social understanding we share with nonhuman animals?

These general arguments indeed provide reason to doubt that human social understanding is purely a matter of folk psychology. The subpersonal processes enabling our navigation of the social world is surely much more complicated than suggested by the traditional folk psychological picture, involving nonmentalistic modes of social understanding such situational understanding and behavior models.

But these general arguments are far from establishing the phenomenological critics' radical claim that folk psychology is restricted to offline, reflective cognition, and plays no role in online social interaction. The mere existence of alternatives to folk psychology does not prove that folk psychology does not play a pervasive role in online social understanding. It seems we would need a careful ethology of human social navigation, a survey of the social navigation problems humans actually confront, in order to determine whether these nonmentalistic alternatives would suffice for the majority of human social interactions. Even then, it must be recognized that the same task of social navigation may be accomplished in a variety of ways. While folk psychology may seem complicated in comparison to situational understanding and other forms of nonmentalistic social understanding, we actually know surprisingly little about the cognitive psychology of folk psychology. We generally lack processing models of how we engage in folk psychological reasoning in the real world. And we know almost nothing about how cognitively taxing belief–desire reasoning is for humans. So it is premature to conclude that folk psychological reasoning is cognitively complex and thus occurs less often than nonmentalistic modes of social understanding.

The critics are right, however, to point out that the standard empirical evidence offered in favor of the folk psychological picture primarily concerns offline rather than online social understanding. We should be careful not to make the mistake of simply assuming folk psychology drives online social understanding. To avoid this mistake, I will focus on an aspect of social understanding that is widely recognized to *require* folk psychology, and cannot be accomplished by nonmentalistic means—namely, performance on standard third-person false-belief tasks. In the next section, I will

explain why false-belief understanding is seen as necessary for successfully navigating such tasks. Then, in section 5, I will provide I will provide empirical evidence that we can indeed use false-belief understanding for online purposes. This is exactly what is needed to establish a role for belief–desire psychology in online social interaction.

4. Why False-Belief Tasks?

If performance on false-belief tasks is going to provide evidence about the pervasiveness of folk psychology, we must be certain that successful performance on such tasks indeed requires folk psychology. Above I stated that it is widely recognized that folk psychology is required to navigate false-belief tasks. But why is this? Here I will explain why false-belief tasks are treated as tests of genuine mental state understanding, and how false-belief tasks requiring online responses could provide evidence against the phenomenological critics.

While there are various types of false-belief tasks now found in the literature, what I have been calling the "standard false-belief task" is the third-person, change-of-location false-belief task. This task was first described independently by Bennett (1978), Dennett (1978), and Harman (1978) in their commentaries on Premack and Woodruff (1978) as a way of testing whether nonhuman primates actually possess a "theory of mind." Wimmer and Perner (1983) were the first to use this task experimentally to test young children's understanding of false belief. As I did in chapter 1, I will use Baron-Cohen et al.'s (1985) version involving characters named Sally and Anne. As a participant in this task, I observe Sally put a marble in the

basket, then while she is gone, Anne move the marble to a box. Focusing on the task of prediction (rather than explanation, since explanation, at least as it is operationalized in the experimental literature, seems essentially tied to offline, reflective cognition), my goal as a participant is to predict where Sally will look for her marble when she returns to the room. According to belief–desire psychology, people's actions on the environment are driven by what they believe about the state of the world, and what goals or desires they have, i.e., what states of the world they want to bring about. In the false-belief task, I am informed that Sally wants to find her marble (e.g., by Sally indicating it in her verbal or nonverbal behavior, or the experimenter simply telling us this). Given this information about her goal, I need to figure out where Sally believes her marble is located in order to predict where she will look for it.

In order to understand why passing the false-belief task requires appreciating beliefs as representational mental states, imagine another scenario, where Sally knows the marble's true location (in the box). Wouldn't predicting her behavior in this case be evidence of possessing a belief–desire psychology, i.e., wouldn't it require an understanding of Sally's representations of the world (specifically, where she believes the marble to be located) and her representation of a desired state of affairs (possessing the marble)? The standard line is that this kind of behavioral prediction could be accomplished without any understanding of representational mental states, i.e., without full-fledged folk psychology.

Consider first Sally's desire for the marble. Our mature concept of DESIRE is usually characterized, like BELIEF, as representational—that when a person desires

something, it means they possess a mental representation whose content describes how they would like the world to be. But Bartsch and Wellman (1995) argue there is a simpler, nonrepresentational notion of desire which children possess before the representational understanding of desire. On their account, a "simple desire" is a mere relation between a person and an external object. This relation can be characterized as *intentional* because the person's simple desire is "about" a particular object in the world. But a simple desire is not *representational* because it does not require "anything like an internal cognitive representation of an object" (p. 13). Gergely and Csibra (2003) for this reason call such an understanding of desires or goals "nonmentalistic." This nonmentalistic understanding of desire seems sufficient to appreciate what Sally wants (since the object of her desire is a particular external object). But to make a prediction about how Sally will behave in order to obtain her marble, I also need to address the belief component. Simply knowing what someone wants, without also knowing anything about the object of their desire (e.g., where it is located), is insufficient to form a prediction how they will behave. Because in this scenario Sally has only *true* beliefs, we can also handle these in a nonmentalistic, nonrepresentational fashion. Instead of appreciating how Sally represents the world as being, I can simply take into account *what I believe about the world*. Thus, it seems as if I can predict the behavior of someone with only true beliefs without possessing a true belief–desire psychology if I make use of (a) a nonmentalistic, nonrepresentational understanding of desire, and (b) what I myself believe about the world.

This will not work, however, in the case of *false* beliefs. To successfully predict how someone with a false belief will behave, I cannot simply attend to the current state of the world. I must understand that they are representing the world as being a certain way, that this representation fails to match how the world actually is (or at least how *I* take it to be), and make my behavioral prediction based on their false representation of the world. In the case of Sally, simply knowing where the marble is actually located will not give me reliable information about where she will look for it. I must be able to track where she represents the marble as being located—i.e., I must know that at a previous point in time she had perceptual access to the marble, which produced in her a belief about its location, and that this belief has not changed as the marble's location has changed. In the case of the standard change-of-location false-belief task, it is likely that an extensional, nonrepresentational understanding of desire would be sufficient to understand that Sally wants the marble: instead of treating her as representing a hypothetical state of affairs, we could treat her as simply being related to an actual object in the environment. But without an understanding of Sally's belief, this nonmentalistic understanding of her desire is inadequate to generate an accurate prediction of her behavior.

Therefore, change-of-location false-belief tasks are treated as evidence of genuine mental state understanding because to predict the behavior of people with false beliefs requires an understanding of their mental representations of the world. While others' behavior that is driven by true beliefs can be predicted by applying belief–desire psychology, there are nonmentalistic means for making the same

predictions. In short, the standard line is that if we want to be sure that a social agent is using folk psychology, false-belief tasks are the way to go.

Admittedly, false-belief understanding has often been given undue importance in the folk psychology camp—to such an extent that it has sometimes been treated as synonymous with acquiring competence in folk psychology (see Bloom & German, 2000). By choosing to focus here on false-belief understanding, I do not endorse its problematic use in philosophical and empirical strands of the folk psychology literature. Of course it must be recognized that other forms of mental state understanding are developmentally more fundamental, and that the development of folk psychology does not end when children can pass standard false-belief tasks around age 4. Further, we do not have much empirical evidence at this point about exactly how important false-belief understanding is to our daily lives, or how easy or difficult it is even for adults (I will say more about this below). But acknowledging all of this does not detract from false-belief understanding's status as a paradigm case of mental state understanding.

Given this status, if it can be shown that people display an understanding of others' false beliefs in online contexts, rather than in cases of reflective, offline cognition, it would serve as strong evidence against the phenomenological critics' skepticism about the role of folk psychology in unreflective, online social interaction. While it would not establish *how often* we use folk psychology in online interactions, it would still cut against the phenomenologists' claim that belief–desire psychology plays *no role* in online social understanding. And while more careful, empirical study would be needed, if these cases of false-belief understanding are like ones we confront

in our everyday lives, it would suggest that belief–desire psychology might play a *substantial* role in online social understanding. If false-belief understanding is the most difficult aspect of belief–desire psychology, and we can establish the use of false-belief understanding in online contexts, we might very well use folk psychology in the more common cases where others' beliefs are true rather than false, even if folk psychology is not *required* in these cases. That is, if we can indeed use false-belief understanding in online texts, it is more parsimonious to think folk psychology plays a pervasive role in online social understanding, than to think it is restricted to the online understanding of others' false beliefs and not used when others have true beliefs. Below I will provide empirical evidence that false-belief understanding can indeed be expressed in our online responses, thus arguing that the phenomenological critics have given too limited a role to mental state understanding in our everyday, unreflective, online social understanding.

5. Evidence for Online False-Belief Understanding

In this section I will describe several recent behavioral experiments on false-belief understanding in children and adults, which can be characterized as testing online rather than offline social understanding. I'll start with two nonverbal tasks used with children as young as 15 months old (Onishi & Baillargeon, 2005; Southgate, Senju, & Csibra, 2007). Rather than require subjects to verbally or nonverbally make explicit predictions of behavior based on attributions of false beliefs, these experiments test other ways in which false-belief understanding may manifest itself in

children's online behavior, specifically their looking behavior. Just as in the Sally–Anne task, these experiments involve false beliefs created by a change in the location of an object.

Onishi and Baillargeon (2005) used a violation-of-expectation paradigm to test whether 15-month-old infants have at least a rudimentary understanding of others' false beliefs. After being familiarized with the scene of an agent hiding a toy in one of two locations, then returning later to retrieve the object from that location, infants were shown the toy's being moved without the agent's knowledge. Infants were then presented with the agent searching for the hidden toy either (a) where the agent falsely believed it to be, or (b) where it was actually located. Infants reliably looked longer at instances of (b), the so-called "unexpected" event, assuming the child expects the agent to search for the toy where she believes it to be located. This experiment tests children's online understanding of others' false beliefs—i.e., children's unreflective expectations about people's behavior given what children know about their epistemic states—rather than children's ability to verbally or nonverbally *report* these expectations to a questioner (or even, seemingly, to themselves).

A problem with looking-time experiments is that they are open to many interpretations about why infants look longer at one condition versus another. As Southgate et al. (2007) note, Onishi and Baillargeon's infants might implicitly attribute *ignorance* to the agent rather than a *false belief*. Thus, infants might look longer at the incongruent event (where the agent acts contrary to her false belief) because they do not expect an agent ignorant of an object's actual location to search for it at that location, rather than because they expect an agent to search for an object

in the location she falsely believes it to be located. Southgate et al. attempted to disambiguate these possibilities using a predictive looking paradigm, where they measured infants' anticipatory eye movements prior to seeing an agent searching for a hidden object. Just as in the previous study, infants were first familiarized with video of an agent watching a toy being hidden in one of two boxes, pausing for a short delay, and then reaching for the toy in that box. In test trials, after the toy was hidden, it was removed from that box while the agent was still not looking. This was done to prevent children from being biased in their looking behavior by knowing the actual location of the toy. The agent next returned to looking at the two boxes, paused for a short delay, then reached for the toy in one of the two locations: where she believed it to be located (where she saw it hidden), or in the other box where she would have no reason to expect it to be hidden. Using eye-tracking technology, experimenters examined where children first looked after the delay. This served as a measure of where the child expected the agent to search for the toy. Before the agent reached for one of the two boxes, these 25-month-olds more often made their first looks toward, and spent more time looking at, the location in accord with the agent's false belief. Thus, their looking behavior suggests that the infants expected the agent to look for the toy where she falsely believed it to be located.

Admittedly, these studies do not fully fit the paradigm of "online" social understanding, since they require children only to passively observe another's

behavior rather than to actually interact with them.[28] Nonetheless, the understanding of false belief required of children seems well characterized as sensorimotor (involving bodily responses to observable stimuli), implicit (not requiring conscious thought) and spontaneous (not requiring explicit instruction from experimenters). It is thus clearly much closer to the "online" end of the cognitive spectrum than the "offline" end, where we find standard false-belief tasks. Furthermore, it is easy to imagine how the implicit understanding displayed in children's looking behavior could be extended to cases of actual social interaction. If you can anticipate where a friend with a false belief will look for a desired object, you might help them out by verbally or nonverbally informing them of the object's actual location. While the young children in the above studies may not yet be able to make use of their false-belief understanding in this way, such a response would be of the same general kind as that displayed in these studies, and clearly meet all the criteria for online social understanding.

Other recent studies of false-belief understanding in older children and adults (Carpenter, Call, & Tomasello, 2002; Keysar et al., 2003) focus on actual social interactions, and thus serve as examples of full-fledged online false-belief understanding. In these behavioral tasks, participants interpret the speech of an interactive partner in light of their false belief about some feature of the task environment. Importantly, no offline reflection about their partner's mental states is

[28] I want to thank an anonymous reviewer for the *Journal of Consciousness Studies* for pressing me on this point, and for highlighting conversation as an example of online mental state understanding found in many traditional folk psychological accounts.

required. Rather, participants must respond online to their interactive partner in a way that requires false-belief understanding.

Consider first the task given to 3-year-olds by Carpenter et al. (2002). Two experimenters (E1 and E2) gave the child two novel objects (A and B) to play with, then taught the child to play a "hiding" game with them. E1 acted as the "hider," placing the target object (A) in a container and the nontarget object (B) on the floor to the side of the container, then closing the container. E2 played the "retriever" role, taking the objects from the container and floor and placing them back in front of the child. In the false-belief condition, E2 left the room, then E1 switched the objects' locations, putting B in the container and hiding A in her bag. E2 returned and tried to retrieve the object in the container, repeatedly using a novel word such as "toma" to name it—saying things like "I'm going to get the toma and then we can play with it." E2 was unsuccessful in opening the container, so E2 and the child instead played with another toy across the room. During this time, E1 placed objects A and B next to each other on a chair. The child was then presented with a retrieval task: E2 noticed the objects, and asked the child to retrieve the object named by the novel word, saying, e.g., "Oh, look, there's the toma! Can you go get the toma and we'll play with it over here." To succeed, the child must understand that the novel word names the object E2 falsely believes to be hidden in the container—namely, object A— and use this information to bring E2 the appropriate object. No reflective judgment (e.g., an explicit report of E2's false belief) is asked of the child; instead the child must respond online to E2's request by retrieving the correct toy.

In this study, the child-participant directly interacts with a person holding a false belief, and must understand that person's false belief to successfully negotiate the interaction. Keysar, Lin, and Barr's (2003) study with adults similarly tested the online use of false-belief understanding during verbally-mediated social interaction, using a modified version of Keysar, Barr, Balin and Brauner's (2000) "referential communication game."[29] In Keysar et al.'s (2000) version, participants sat on one side of a grid containing various objects, with a confederate on the other side playing the role of "director," instructing the participant where to move objects around the grid. While some of the objects were mutually visible to the participant and the director, others were visible only to the participant. Accordingly, some of the director's instructions were designed by the experimenters to be ambiguous from the participant's perspective, but not from the director's perspective. For example, the director and participant could both see a three-inch-high candle and a two-inch-high candle, but a one-inch-high candle was also visible only to the participant. Thus when the director said to "Move the small candle to the right," the participant would need to take into account the director's visual perceptive and knowledge to know that "the small candle" referred to the two-inch candle (the smallest candle from the director's perspective) rather than the one-inch candle (the smallest from the participant's perspective).

[29] Keysar et al. (2003) explicitly describe their task as testing the "spontaneous, non-reflective use" of folk psychology, as opposed to traditional tasks that test how we use it "reflectively and deliberately" (p. 28). Dumontheil, Apperly, and Blakemore (2010) explicitly use my preferred language, describing this task as an "online communication game" (p. 332) testing the "online usage of theory of mind" (p. 331).

Keysar et al. (2003) modified the communication game to test adults' ability to appreciate the director's false beliefs about objects in the grid. They did so by having the participant (out of sight of the director) hide one of the objects—e.g., a roll of tape—in a paper bag and place it in a spot on the grid not visible to the director. The experimenter then misinformed the director about the contents of the bag—e.g., indicating it contained a small ball rather than a roll of tape. The director would then give the participant an instruction that was ambiguous from the participant's perspective but not the director's—e.g., "Move the tape," when a cassette tape was mutually visible. To determine whether the director's instruction referred to the hidden roll of tape or the cassette tape, participants needed to know that the director had a false belief about the hidden contents of the bag, and thus could not be referring to the roll of tape with the word "tape." Participants' understanding of the director's instructions were measured by what object they first looked at and then reached for.[30]

The experimental tasks in Carpenter et al. (2002) and Keysar et al. (2003) provide evidence of social interactions which require online responses to people's false beliefs. While it is hard to gauge the ecological validity of such experimental situations, they are clear cases where correctly interpreting the speech of an interactive partner requires appreciating their false beliefs about an object relevant to their interaction. Since these tasks involve verbally-mediated social interaction, they undoubtedly fit the paradigm of online social understanding. Participants are not being

[30] Keysar et al. (2003) and other related studies show that adults, and not just children, often fail to take into account another person's beliefs when they diverge from their own. In Keysar et al. (2003), participants reached for the hidden object on 22% of false-belief trials.

asked to make reflective judgments about the other person's mental states, or to explicitly predict or explain their behavior. Rather, false-belief understanding is required to successfully navigate the interaction, to respond online to the other person's verbal request.

In summary, the nonverbal and verbal experimental tasks I have described serve as cases of *online* false-belief understanding because they measure unreflective, spontaneous responses to another person possessing a false belief. While these studies do not address the phenomenological experience of being in these situations, they do not require participants to engage in offline reflection about others' false beliefs. Rather, their false-belief understanding is demonstrated in their online behavior.

Before moving to what the phenomenological critics might say about these cases, I will more explicitly describe how folk psychological accounts would characterize the psychological processes driving these online responses. Whether they simulate others' mental states or apply theoretical knowledge, the standard folk psychological picture is that people possess a concept of BELIEF, and are able to make false-belief ascriptions based on their perception of others' behavior; such false-belief ascriptions are what cause the online responses measured in these studies, as well as the offline, reflective judgments in standard false-belief tasks. For example, one could interpret the infants in the Southgate et al. (2007) study as (implicitly) thinking the following: *that person believes the toy is in the box on the left*, *and he intends to reach for the toy*, *so I predict he will reach into the box on the left*. These children cannot yet articulate this knowledge verbally, but, on this view, the same knowledge about mental states required for the Sally–Anne task is present in these 1.5- to 2-year-old

infants. There is no deficit in children's conceptual knowledge of belief at this young age; that this knowledge manifests itself in their looking behavior but not other kinds of behavioral responses is attributable to performance deficits.

The verbal studies are a bit more complicated to characterize since language is involved. But the standard folk psychological account is that language comprehension and production essentially involve mental state attribution: that we interpret people's utterances by inferring their intended meaning, and consider the mental perspective of our audience when speaking (e.g., Sperber & Wilson, 1986/1995, 2002; Tomasello, 2003). In Carpenter et al. (2002), for example, the child must recognize that the experimenter falsely believes that the target object is in the container. This is the only way the child could interpret the novel word uttered by the experimenter when attempting to open the container as referring to the target object, rather than to the non-target object actually located inside the container. According to the folk psychological account, such understanding again crucially depends on conceptually representing the other person as possessing a false belief.

6. Possible Responses from the Phenomenological Critics

The phenomenological critics would surely object to such folk psychological characterizations of online social understanding. It goes against everything they say about the minor role of folk psychology in human social understanding to admit that it can drive online social understanding. It is precisely by relegating folk psychology to offline social understanding that the phenomenological critics argue that it is less

important to our everyday navigation of the social world than is traditionally assumed (again, on the assumption that online phenomena are more common than offline phenomena). And for them to accept the folk psychological account of these studies, but to say that online false-belief understanding is an exceptional case and that the rest of folk psychology operates only in offline cases, would be entirely ad hoc. Why would we attribute beliefs online when others have *false* beliefs, but not in the more common cases when they have *true* beliefs? False-belief understanding is generally treated as one of the most difficult aspects of our mature folk psychological abilities. Why would the more difficult form be the one we can do spontaneously in the course of online interaction?[31]

So to defend their critique of folk psychology, the phenomenological critics would need to reject the folk psychological interpretation of these studies as involving online false-belief understanding. But what precisely would they reject about this interpretation, and why? Even though they claim that we can directly perceive emotions and intentions in people's behavior, the critics surely cannot treat people's false beliefs in the same way. What is interesting about false beliefs is that they are not currently perceivable, and thus paradigmatic of why we treat mental states as "inner," "hidden" and distinct from observable behavior. To understand false beliefs, we must understand that people have points of view on the world which can fail to accord with the world's actual state. It seems that the people in the above studies are indeed

[31] I would like to thank Sam Rickless for pressing me to develop this point.

responsive to others' *beliefs*—i.e., inner states of epistemic agents serving to represent the world. Situational understanding or other similarly nonmentalistic alternatives to folk psychology would also be inadequate to handle cases involving false beliefs. Situational understanding, for example, is a matter of understanding how people *normally* act in particular situations with particular roles. But surely our understanding of, say, the waiter role assumes the waiter has true rather than false beliefs (about, e.g., what is on the menu, where things are located in the restaurant). To take into account patterns of behavior when people have true versus false beliefs would require tracking when people do indeed have these kinds of mental states, and thus become a form of folk psychology. The alternative forms of social understanding offered by the phenomenologists and other researchers thus seem no help in understanding behavior driven by false beliefs—something humans are clearly able to do. So what should be said on their behalf about false-belief understanding and its apparent use in online contexts?

One alternative interpretation open to the critics is that people's online behavior is not actually responsive to other people's false beliefs, but to other properties which are often correlated with their beliefs. Ratcliffe (2007, pp. 53–54), for example, suggests people could solve standard change-of-location false-belief tasks by following a behavior rule that people look for things where they last saw them. Skeptical challenges like this have repeatedly been put forward against purported behavioral evidence of mental state understanding in nonhuman animals or young children. All researchers must of course respect appeals to parsimony. But these

are claims that can be empirically tested by designing experimental tasks where simpler, nonmentalistic methods break down.

For example, Josef Perner and Ted Ruffman (Perner & Ruffman, 2005; Ruffman & Perner, 2005) have expressed just such a skeptical interpretation of Onishi and Baillargeon's (2005) study. They pose a few possible alternative explanations of the infants' looking behavior in this study which fall short of true false-belief understanding, including the idea that infants use learned behavioral rules, such as the "people look for objects where they last saw them" rule Ratcliffe describes. But they explicitly acknowledge what would constitute evidence that infants possess true false-belief understanding as opposed to simply using behavioral rules: that they perform successfully on a variety of tasks "in which behavior rules would lead to contradictory predictions of actions"(Perner & Ruffman, 2005, p. 216).[32] Penn and Povinelli (2007)

[32] Ruffman and colleagues (Garnham & Ruffman, 2001) used just such a method in previous work testing the false-belief understanding of somewhat older children (3- to 4-year-olds). Following Clements & Perner (Clements & Perner, 1994), Ruffman et al. tested whether older children display an "implicit" appreciation of others' false beliefs in their anticipatory looking behavior. Unlike the more recent nonverbal studies I described above, these studies used the verbal method of standard false-belief tasks, where an experimenter verbally narrates and acts out the change-of-location scenario using toy figures. The novel feature of these older studies as compared to standard false-belief tasks was to measure children's looking behavior when the experimenter wonders aloud about the protagonist of the story, "I wonder where he's going to look." Garnham and Ruffman (2001) designed a version of the standard change-of-location task involving three rather than two possible target locations, to test, amongst other things, whether children's looking behavior is driven by a "seeing = knowing" rule. According to this rule, if an agent sees the location of an object, he knows where it is, and thus will be successful in searching for it in the future. Conversely, if an agent has not seen where an object is actually located, as in a false-belief condition of a change-of-location false-belief task, children applying this rule would reason that the agent must not know where it is located, and thus will do the wrong thing by searching where the object is not. In a change-of-location task with three locations (left, middle, and right), if the target object is moved from the left-hand location to the right-hand location, the "seeing = knowing" rule predicts children's anticipatory looking will be directed to either of the two incorrect locations (either the left-hand or middle locations). But if children understand false beliefs, they will look preferentially toward the left-hand location, where the protagonist falsely believes the object is, but not toward the middle location. Garnham and Ruffman found that 3- to 4-year-olds looked

articulate this same idea in more formal terms when discussing what would constitute convincing experimental evidence that nonhuman animals can understand mental states. Without getting into further detail about experimental methodology or particular studies, my point is that nonverbal behavior is widely acknowledged as a potential indicator of mental state understanding. The phenomenological critics would not be on very firm ground to suggest that simpler, nonmentalistic methods can account for *all* purported cases of online false-belief understanding, including the online behavior of adults. Admittedly, the studies with very young infants require further study before any definitive conclusions can be made about when we acquire genuine false-belief understanding. But when these nonverbal studies are combined with the tasks requiring verbally-mediated social interaction, it is unclear how simple behavioral rules could account explain account for any individual's successful navigation of all of them. The situational contexts and response types described in the studies above are too varied for this general skepticism to hold weight (cf. Call & Tomasello, 2008).

A more substantial objection by Gallagher (2005) and Ratcliffe (2007, pp. 205–211) offers a specific alternative to folk psychology's account of beliefs as inner, representational states: the view that beliefs are "dispositions to act and to experience in various ways" (2005, p. 214). Both authors suggest that having a belief does not

---

more to the location where the protagonist falsely believed the object to be located, than to either of the other two locations, and that there was no significant difference in looking time between these two incorrect locations. This behavior is inconsistent with the application of the "seeing = knowing" rule.

involve possessing a discrete internal state, but rather that belief attributions can be indeterminate and "ambiguous even from the perspective of the believer" (2005, p. 215). It is not clear from these authors' writings, however, what a dispositional theory of belief is a theory of. Is it an account of what beliefs really are? Is it an account of what everyday people take talk about beliefs to be referring to? Is it an account of what we represent about other people's epistemic states for the purpose of online behavior? Ratcliffe's (2007, ch. 7) discussion of belief is mostly about the wide range of uses for the term "belief" in everyday discourse. Gallagher's (2005) discussion wavers, sometimes referring to what we think and talk about other people in understanding their verbal and nonverbal behavior, and sometimes referring to whether a person "in reality" has a particular belief.

I find such a dispositional account of belief at least relevant to characterizing our *talk* about beliefs—I doubt this discourse is as simple as standard folk psychological accounts suggest. But I am less satisfied with such an account when attempting to explain the psychological processes by which we track people's epistemic states and act in light of such understanding—i.e., when the focus is on online social understanding. In the tasks described above, we're considering very discrete epistemic states of agents: where they believe a particular object to be located, or what they believe to be found at a particular location. Why not treat these as representational states of agents, and my understanding of these representational states as (meta-)representing them? As discussed above, treating agents as tracking behavioral dispositions is an alternative offered by researchers skeptical of attributing mentalistic understanding. So the burden is on the phenomenological critics to show

how an account of beliefs as complex behavioral dispositions would differ from nonmentalistic representations of behavior, and why we should call one mentalistic and the other nonmentalistic, if both types of account are merely concerned with patterns of and dispositions for behavior.

The discreteness of the mental states in question also addresses a related objection offered by Ratcliffe (2007, ch. 7): that although researchers in the folk psychology camp describe various situations as involving "belief understanding," there is not actually a unitary phenomenon deserving this name; people can appreciate a variety of psychological features of other people, so it is unclear what exactly is being tested by experiments of "belief understanding." The psychological phenomena Ratcliffe thinks we can distinguish but which are mistakenly lumped together by the folk psychological account of "belief understanding" include: sentential attitudes (attitudes directed toward sentences, of the form "A believes that the sentence 'S' is true") versus propositional attitudes (attitudes directed toward states of affairs that can be expressed as propositions, of the form "A believes that p"); behavior driven by explicit thought versus habitual behavior involving no such explicit thoughts; and commitments and convictions that shape our experience, attitudes, and actions in a way distinct from the psychological profile of mere propositional attitudes (e.g., a "belief" in the existence of God). From such cases, Ratcliffe concludes that there is no unitary concept of "belief," and that the folk psychological account of belief–desire psychology is a misleading characterization of human social understanding.

I agree with Ratcliffe that our understanding of people's behavior is not simply a matter of attributing beliefs and desires, i.e., propositional attitudes playing

(respectively) informational and motivational roles. This surely oversimplifies the nature of mental state understanding, in both its online and offline forms. But I do not accept Ratcliffe's conclusion that the folk psychological picture is so oversimplified as to be false, that there are no unitary concepts of belief and desire playing a role in actual human social understanding. More specific to my argument here, whether or not people in the folk psychology camp have overextended the term "belief" does not affect the interpretation of the cases of online false-belief understanding I've described. As argued above, these studies concern a well-defined phenomenon: being sensitive to people's discrete beliefs about particular objects at particular locations. None of the distinctions Ratcliffe makes call into question the unity of this phenomenon, or the folk psychological account of this phenomenon in terms of appreciating other people's false representations of the world. How the understanding displayed in these studies relates to other forms of social understanding described in terms of "belief understanding" is an open question. But these experiments are representative of how false-belief understanding is studied experimentally, and show how it can mediate our online social interactions.

Another interpretation open to the phenomenological critics involves the idea that "tracking" false beliefs for the purposes of online behavior does not depend on conceptually representing them.[33] But when presented in this negative form, this alternative need not stray very far from the folk psychological account. For example,

[33] I want to thank another anonymous referee for the *Journal of Consciousness Studies* for calling attention to this possible response.

following up on a discussion of Robert Gordon's version of ST, Dokic (2002) describes a simulation-based, nonconceptual understanding of belief capable of driving online behavior. On this view, a concept of BELIEF is required to use the product of a mental simulation to have thoughts or make utterances ascribing a belief to a person. But without the concept BELIEF, a person could still use the results of a simulation routine—i.e., the information about the other's beliefs gained from pretending to have those beliefs (where "pretending" is not necessarily conscious or explicit)—to drive their behavior. Such a person shouldn't be said to be making unconscious or implicit belief ascriptions, as the person never entertains thoughts using the concept BELIEF, as required by the standard folk psychological account. Yet the person is indeed using information about the person's mental states gained from a simulation process. It is possible that online false-belief understanding is driven by such nonconceptual simulation processes, while offline false-belief ascriptions involve conceptual representations, as described by the standard folk psychological account. Another possibility is that nonconceptual simulation characterizes the immature false-belief understanding found in very young children, while adult online false-belief understanding is driven by a conceptual understanding of belief. Either option involves a departure from the standard folk psychological account that is open to the phenomenological critics. But both retain the core of the folk psychological account, that we entertain representations of others' mental states.

A more radical account along these lines is that online belief tracking does not involve representing other's beliefs at all. Hutto (2008a), for example, develops a nonrepresentational "biosemiotic" account of the online tracking of others'

psychological states. I do not have space here to go into the details of Hutto's view, but one point is especially significant. According to Hutto, language is required to represent the intensional content of propositional attitudes—i.e., that the same object can be represented in different ways or under different descriptions. Accordingly, Hutto believes we can only nonverbally track people's "intentional attitudes"—i.e., their intentional relations to states of affairs which are nonrepresentational, noncontentful, and extensional (as opposed to intensional) in nature. Intentional attitudes can, for Hutto, be evaluated in terms of their "success" or "error," but not in terms of their truth, as is the case with propositional attitudes. Although he does not explicitly address the purported cases of online false-belief understanding I've described above, Hutto's account would likely characterize them as responding online to people's extensionally "misaligned" intentional attitudes, rather than to false beliefs understood as propositional attitudes. One reason in favor of this interpretation is that these online false-belief tasks (with the exception of Keysar et al., 2003[34]) and standard false-belief tasks arguably do not require attending to the intensionality of other people's beliefs (see, e.g., Apperly & Robinson, 2003). They require understanding that another's beliefs can be extensionally off target (e.g., believing that an object is located somewhere it isn't actually located), but not the referential opacity of their beliefs, i.e., that they represent objects under certain descriptions but not

[34] Understanding the intensional nature of belief is required for some of the conditions in Keysar et al.'s (2003) study—e.g., appreciating that a particular object is well described as "the small candle" from the director's perspective but as "the middle-sized candle" from the participant's perspective. But not all of their conditions required participants to understand that beliefs can represent objects under particular descriptions.

others (e.g., that some object is a green ball, but not that it is Sally's favorite toy). Hutto's contrast between intentional and propositional attitudes respects these different aspects of belief understanding. The more controversial part of Hutto's account is his characterization of intentional attitudes as nonrepresentational, noncontentful, and not truth-evaluable, and our understanding of others' intentional attitudes as also exhibiting these properties. We might plausibly deny these claims, and reject Hutto's biosemiotic account of how we are nonrepresentationally sensitive to such states. This would lead us back to the two options we had before: (a) a unified account of beliefs as representational states of persons and a unified account of how we understand them, as in the standard folk psychological account, or (b) a combination of the conceptual, folk psychological account with a nonconceptual simulation-based account. But even if we accept Hutto's account of nonrepresentational intentional-attitude tracking as characterizing some purported cases of online false-belief understanding, we need not accept that it covers *all* our online responses. It would certainly be more parsimonious if all online responses involved nonrepresentational tracking of intentional attitudes, leaving an understanding of belief-qua-propositional-attitude to offline reflection. But no convincing argument has been offered that this is case, that propositional-attitude understanding *cannot* drive online responses.

Properly evaluating Hutto's account of nonrepresentational intentional-attitude tracking is beyond the scope of this chapter. At this point, it is an avenue which the phenomenological critics could pursue to draw a wedge between online responsiveness to and offline reflection about false beliefs. This interpretation would,

however, require the phenomenological critics to concede my point that online behavior can indeed be driven by *mentalistic* understanding—even if the kinds of mental states at issue are not fully fledged propositional attitudes, as the folk psychological account contends.

In more recent writings, Gallagher and Zahavi have gone beyond their appeal to nonmentalistic social understanding, directly addressing false-belief and other forms of mental state understanding (Gallagher & Zahavi, 2008; Gallagher & Hutto, 2008). Continuing their critique of standard accounts, they reject folk psychological theorizing and simulation as the basis for such understanding. They instead appeal to Hutto's (2008a) account of folk psychological understanding as a *narrative* competency which we develop by engaging with others in story-telling practices about people's reasons for action. Given its very recent introduction to the debate, the narrative account has yet to be fully vetted as a genuine alternative to TT and ST. It is important to note, however, that these phenomenological critics continue to treat folk psychological understanding as only necessary in "puzzling" cases where other nonmentalistic modes of understanding break down (e.g., Gallagher & Zahavi, 2008, p. 193). Yet they have largely left unspecified exactly what counts as a "puzzling" piece of behavior. According to the phenomenological critics, is behavior driven by false beliefs necessarily experienced as "puzzling"? The online false-belief experiments described above were specifically constructed so as to require an appreciation of another person's false beliefs. But they only measured participants' behavior, so they do not provide evidence about participants' conscious experience when in such situations. My hypothesis, however, is that at least some of the time we

can respond online to people's false beliefs without experiencing their behavior as puzzling.[35] The burden is on the phenomenological critics to provide an argument that this is not possible. Therefore, even if Hutto's narrative account offers a viable alternative to TT and ST, nothing about it precludes narrative understanding from driving online behavior in addition to offline reflection.

In summary, I have surveyed several alternatives to the folk psychological account of false-belief understanding open to the phenomenological critics. None, however, have offered persuasive reason to deny that false-belief understanding is driving the online responses described in section 5. These experiments thus provide reason to reject Gallagher's (2005) assertion that "The science of false-belief tests does not provide any evidence for the claim that theory of mind processes are implicit or subpersonal" (p. 219). Admittedly, Hutto's nonrepresentational account of intentional-attitude tracking may be a viable way of treating these online responses as involving less than a full-fledged understanding of belief-qua-propositional-attitude. But even so, intentional-attitude tracking is a form of *online mentalistic* understanding, and thus serves as evidence against the phenomenological critics' claim that online social understanding is purely nonmentalistic. What online responses, in what contexts, are mentalistic or nonmentalistic is an open question—but it is an empirical

[35] Keysar et al.'s (2003) study may speak to this issue. Their behavioral results indicate that our initial, automatic response is to attribute to others what we ourselves believe, and that only through a subsequent correction process can we alter this initial egocentric attribution and represent others' beliefs different from our own. That we often actually reach for the wrong object suggests that at least some of the time we consciously notice this error in order to correct for it. Of course this only suggestive, and research directly studying our conscious experience in such situations will be necessary to settle this issue.

matter for scientists to address. This case study of online false-belief understanding makes the simpler, conceptual point that the phenomenological critics are wrong to rule out online forms of folk psychology.

7. Conclusion

I hope to have exposed a limitation of the phenomenological critics' attack on folk psychology. These experiments of false-belief understanding do support the critics' contention that online intelligence is very important to social understanding, and that many folk psychological accounts have failed to properly acknowledge this. But the critics have yet to make their case that all online social understanding should be couched in nonmentalistic terms.

How to characterize online versus offline intelligence is a general problem raised by embedded, embodied accounts of the mind, including those based in the phenomenological tradition of philosophy. The folk psychological account of social understanding is one of the traditional accounts of cognition that must face up to advances made from this perspective. But, if I am correct, folk psychology need not be seen as a relic of traditional cognitive science and philosophy of mind characterizing only a highly restricted range of the phenomena involved in social understanding. The folk psychological account can be updated to fit within an embedded, embodied approach to social understanding. While mental state understanding may not play the all-encompassing role it has traditionally been assumed to have, it is much more significant than the phenomenological critics allow.

Postscript: The Cognitive Psychology of Folk Psychology

I will end with an important topic for future empirical research highlighted by my discussion of online false-belief understanding: the issue of how automatic and unreflective false-belief understanding and other forms of folk psychological understanding are. My discussion in this chapter has appealed to a rather crude way of carving up online versus offline social understanding in terms of overt bodily behavior. The experimental evidence I have described shows that people can use information about others' false beliefs to negotiate social interactions, rather than simply to make reflective judgments about others' behavior and mental states. Since the phenomenological critics work at the same level of detail, my discussion has been sufficient for the purposes of defending the folk psychological picture against these phenomenological critics. But none of these studies provide direct evidence about how spontaneous or automatic their false-belief attributions are. For all we know, the people in these studies must consciously intend to track the mental states of their interactive partners, and consciously think about these states. Other kinds of methods are needed to determine how unreflective false-belief understanding and other forms of mental state understanding are in our daily lives.

Empirical research has just begun on what Ian Apperly (in press) calls the "cognitive psychology of theory of mind." For example, there is currently mixed evidence about whether belief attribution is something we do automatically in response to the appropriate perceptual stimuli, or whether it is something we must

consciously initiate. In the first study directly on this topic, Apperly, Riggs, Simpson, Samson, and Chiavarino (2006) provided evidence that we do not automatically encode and track people's beliefs. In their experiment, participants were shown videos of nonverbal, change-of-location false-belief situations. On each trial, after the protagonist returned to the scene, participants were probed either about reality, i.e., the actual location of the object (e.g., "It is true that the object is on the right"), or the protagonist's belief (e.g., "She thinks that the object is on the right"), to which they responded "yes" or "no" by pressing the appropriate button. Participants were not, however, explicitly instructed to track the protagonists' mental states. They were only told to track the object's actual location. Apperly et al. argue that if belief reasoning is automatic, participants should be able to attend to the protagonists' belief without overt instruction. This would mean participants at the time of the probe should have tracked both the object's location and the protagonists' belief about the object, and be similarly prepared to answer questions about either. But when the researchers compared reaction times to the reality probes versus the belief probes, they found subjects were significantly slower to respond to belief probes. From this, they infer belief reasoning is not something we do automatically whenever observing human behavior. A bit more precisely, Apperly et al. conclude that participants must have failed to ascribe a particular belief to the protagonist by the time of the belief probe. They remain agnostic about the particular subprocesses that may have been present or missing by the time of the belief probe's appearance. For example, they allow that participants might have automatically inferred a set of candidate belief contents for the protagonist from simply watching the video, but, by the time of the belief probe, had

failed to select one of these candidate beliefs and ascribe it to the protagonist. Thus, while Apperly et al. summarize their study as providing as evidence against belief ascription's being automatic, they are a bit more careful in their discussion about what they have actually shown—namely, that belief ascription is not complete prior to explicitly being probed about it; whether or not the process of belief ascription has already begun is not something these experimental results can speak to.

In a follow up study, Cohen and German (2009) provide evidence that we do indeed automatically form determinate belief ascriptions, but that this information decays extremely rapidly. They argue that Apperly et al.'s results are due to there being a long delay between the events that signal the content of the protagonist's belief, and the belief probe. Cohen and German modified the order of events in their experiment to reduce this delay. In this condition, participants actually responded faster to belief probes than to reality probes, and responded no slower than when overtly instructed to track the protagonist's belief. This study thus suggests we do automatically encode other people's beliefs when we perceive their behavior, but that we must use this information quickly or it will decay and no longer be available for use.

It should be noted that while these two studies address whether we encode beliefs without overtly intending to do so, they are far from the paradigm of online social understanding I've sketched in this chapter. They require participants to passively observe a third-person's behavior, and respond to a linguistic description of that person's belief with a "yes" or "no" button-press, rather. Participants are not interacting with the protagonist in question, so attending to that person's mental states

serves no practical purpose for the agent. But if we can find evidence of automatic belief attribution in such an impoverished social situation, we would have every reason to think this happens in the course of our everyday social interactions, where this information might actually be useful to us.

These two studies are surely not the definitive word on the automaticity of folk psychology. But it is just this type of empirical research which will serve to address the issues raised by the phenomenological critics about how much and how easily we use folk psychology in our everyday lives.

Chapters 1 and 5 contain material reproduced from "False-belief understanding and the phenomenological critics of folk psychology," *Journal of Consciousness Studies*, *15*(12), 33–56. Permission to reproduce this material has been granted by the copyright owner, Imprint Academic. The dissertation author was the primary investigator and author of this paper.

References

Apperly, I. A. (in press). *The cognitive psychology of theory of mind*. Hove, UK: Psychology Press.

Apperly, I. A., Riggs, K. J., Simpson, A., Chiavarino, C., & Samson, D. (2006). Is belief reasoning automatic? *Psychological Science*, *17*(10), 841–844.

Apperly, I. A., & Robinson, E. J. (2003). When can children handle referential opacity? Evidence for systematic variation in 5- and 6- year old children's reasoning abut beliefs and belief reports. *Journal of Experimental Child Psychology*, *85*, 297–311.

Aslin, R. N. (2007). What's in a look? *Developmental Science*, *10*(1), 48–53.

Baker, C. L., Tenenbaum, J. B., & Saxe, R. R. (2006). Bayesian models of human action understanding. *Advances in Neural Information Processing Systems*, *18*, 99–106.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and "theory of mind."* Cambridge, MA: MIT Press.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*(1), 37–46.

Bartsch, K., & Wellman, H. M. (1995). *Children talk about the mind*. New York: Oxford University Press.

Bechtel, W. (1994). Levels of description and explanation in cognitive science. *Minds and Machines: Journal for Artificial Intelligence*, *4*(1), 1–25.

Bechtel, W. (2005). The challenge of characterizing operations in the mechanisms underlying behavior. *Journal of the Experimental Analysis of Behavior*, *84*, 313–325.

Bechtel, W. (2008a). Mechanisms in cognitive psychology: What are the operations? *Philosophy of Science*, *75*, 983–994.

Bechtel, W. (2008b). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. New York: Routledge.

Bechtel, W. (2009). Explanation: Mechanism, modularity, and situated cognition. In P. Robbins, & M. Aydede (Eds.), *The Cambridge handbook of situated cognition* (pp. 155–170). Cambridge, England: Cambridge University Press.

Bechtel, W. (in press). The epistemology of evidence in cognitive neuroscience. In R. Skipper Jr., C. Allen, R. A. Ankeny, C. F. Craver, L. Darden, G. Mikkelson, & R. Richardson (Eds.), *Philosophy and the life sciences: A reader*. Cambridge, MA: MIT Press.

Bechtel, W., & Abrahamsen, A. A. (1993). Connectionism and the future of folk psychology. In S. M. Christensen, & D. R. Turner (Eds.), *Folk psychology and the philosophy of mind* (pp. 340–367). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bechtel, W., & Abrahamsen, A. A. (in press). Dynamic mechanistic explanation: Computational modeling of circadian rhythms as an exemplar for cognitive science. *Studies in History and Philosophy of Science Part A*.

Bennett, J. (1978). Some remarks about concepts. *Behavioral and Brain Sciences*, *1*, 557–560.

Bermúdez, J. L. (1995). Syntax, semantics, and levels of explanation. *Philosophical Quarterly*, *45*(180), 361–367.

Bermúdez, J. L. (2003). The domain of folk psychology. In A. O'Hear (Ed.), *Royal institute of philosophy supplement: Vol. 53. Minds and persons* (pp. 25–48). New York: Cambridge University Press.

Bermúdez, J. L. (2005). *Philosophy of psychology: A contemporary introduction*. London: Routledge.

Blackburn, S. (1992). Theory, observation, and drama. *Mind & Language*, *7*(1–2), 187–203.

Bloom, P., & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, *77*(1), B25–B31.

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, *12*(5), 187–192.

Carpendale, J., & Lewis, C. (2006). *How children develop social understanding*. Malden, MA: Blackwell.

Carpenter, M., Call, J., & Tomasello, M. (2002). A new false belief test for 36-month-olds. *British Journal of Developmental Psychology*, *20*(3), 393–420.

Carruthers, P., & Smith, P. K. (Eds.). (1996). *Theories of theories of mind*. Cambridge, England: Cambridge University Press.

Churchland, P. M. (1989). *A neurocomputational perspective: The nature of mind and the structure of science*. Cambridge, MA: MIT Press.

Clark, A. (1993). *Associative engines: Connectionism, concepts, and representational change*. Cambridge, MA: MIT Press.

Clark, A. (2007). Coupling, emergence, and explanation. In M. Schouten, & H. Looren de Jong (Eds.), *The matter of the mind: Philosophical essays on psychology, neuroscience, and reduction* (pp. 227–248). Malden, MA: Blackwell.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7–19.

Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development*, *9*(4), 377–395.

Cohen, A. S., & German, T. C. (2009). Encoding of others' beliefs without overt instruction. *Cognition*, *11*(3), 356–363.

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford, England: Clarendon Press.

Craver, C. F., & Bechtel, W. (2007). Top-down causation without top-down causes. *Biology and Philosophy*, *22*, 547–563.

Cummins, R. (1996). *Representations, targets and attitudes*. Cambridge, MA: MIT Press.

Davies, M., & Stone, T. (Eds.). (1995a). *Folk psychology: The theory of mind debate*. Oxford, England: Blackwell.

Davies, M., & Stone, T. (Eds.). (1995b). *Mental simulation: Evaluations and applications*. Oxford, England: Blackwell.

De Jaegher, H. (2006). *Social interaction rhythm and participatory sense-making: An embodied, interactional approach to social understanding, with some implications for autism.* Unpublished Ph.D. dissertation, University of Sussex, Brighton, United Kingdom.

De Jaegher, H., & Di Paolo, E. (2007). Participatory sense-making. *Phenomenology and the Cognitive Sciences*, *6*(4), 485–507.

Deák, G. O., Bartlett, M. S., & Jebara, T. (2007). New trends in cognitive science: Integrative approaches to learning and development. *Neurocomputing*, *70*, 2139–2147.

Dennett, D. C. (1969). *Content and consciousness*. New York: Humanities Press.

Dennett, D. C. (1978). Beliefs about beliefs. *Behavioral and Brain Sciences*, *1*, 568–570.

Dennett, D. C. (1981). *Brainstorms: Philosophical essays on mind and psychology* (1st MIT Press ed.). Cambridge, MA: MIT Press.

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Dennett, D. C. (1991a). *Consciousness explained* (1st ed.). Boston: Little, Brown and Co.

Dennett, D. C. (1991b). Real patterns. *Journal of Philosophy*, *88*(1), 27–51.

Dôkic, J. (2002). Reply to Pierre Jacob. In J. Dôkic, & J. Proust (Eds.), *Simulation and knowledge of action* (pp. 111–117). Amsterdam: John Benjamins.

Dreyfus, H. (1972). *What computers can't do.* Cambridge, MA: MIT Press.

Dreyfus, H. (1991). *Being-in-the-world*. Cambridge, MA: MIT Press.

Dumontheil, I., Apperly, I. A., & Blakemore, S. (2010). Online usage of theory of mind continues to develop in late adolescence. *Developmental Science*, *13*(2), 331–338.

Farrer, C., & Frith, C. D. (2002). Experiencing oneself vs. another person as being the cause of an action: the neural correlates of the experience of agency. *Neuroimage*, *15*(3), 596–603.

Fletcher, P. C., Happé, F., Frith, U., & Baker, S. C. (1995). Other minds in the brain: A functional imaging study of "theory of mind" in story comprehension. *Cognition*, *57*(2), 109–128.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. Cambridge, MA: MIT Press.

Frith, U. (2003). *Autism: Explaining the enigma*. Malden, MA: Blackwell.

Gallagher, H. L., Happé, F., Brunswick, N., Fletcher, P. C., Frith, U., & Frith, C. D. (2000). Reading the mind in cartoons and stories: An fMRI study of "theory of the mind" in verbal and nonverbal tasks. *Neuropsychologia*, *38*(1), 11–21.

Gallagher, H. L., Jack, A. I., Roepstorff, A., & Frith, C. D. (2002). Imaging the intentional stance in a competitive game. *NeuroImage*, *16*(3, Part 1), 814–821.

Gallagher, S. (1997). Mutual enlightenment: Recent phenomenology in cognitive science. *Journal of Consciousness Studies*, *4*(3), 195–214.

Gallagher, S. (2001). The practice of mind: Theory, simulation or primary interaction? *Journal of Consciousness Studies*, *8*(5–7), 83–108.

Gallagher, S. (2003). Phenomenology and experimental design: Toward a phenomenologically enlightened experimental science. *Journal of Consciousness Studies*, *10*(9–10), 85–99.

Gallagher, S. (2004). Situational understanding: A Gurwitschian critique of theory of mind. In L. Embree (Ed.), *Gurwitsch's relevancy for cognitive science* (pp. 25–44). The Netherlands: Springer.

Gallagher, S. (2005). *How the body shapes the mind*. Oxford, England: Clarendon Press.

Gallagher, S. (2007). Simulation trouble. *Social Neuroscience*, *2*(3–4), 353–365.

Gallagher, S. (2008a). Are minimal representations still representations? *International Journal of Philosophical Studies*, *16*(3), 351–369.

Gallagher, S. (2008b). Direct perception in the intersubjective context. *Consciousness and Cognition*, *17*(2), 535–43.

Gallagher, S., & Hutto, D. D. (2008). Understanding others through primary interaction and narrative practice. In J. Zlatev, T. P. Racine, C. Sinha, & E. Itkonen (Eds.), *The shared mind: Perspectives on intersubjectivity* (pp. 17–38). Amsterdam: John Benjamins.

Gallagher, S., & Zahavi, D. (2008). *The phenomenological mind: An introduction to philosophy of mind and cognitive science*. London, England: Routledge.

Garnham, W. A., & Ruffman, T. (2001). Doesn't see, doesn't know: Is anticipatory looking really related to understanding of belief? *Developmental Science*, *4*(1), 94–100.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naïve theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Goldman, A. I. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford, England: Oxford University Press.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Cambridge, MA: The MIT Press.

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind & Language*, *7*(1–2), 145–171.

Greenwood, J. D. (1999). Simulation, theory-theory and cognitive penetration: No "instance of the fingerpost." *Mind & Language*, *14*(1), 32–56.

Grush, R. (2003). In defense of some "Cartesian" assumptions concerning the brain and its operation. *Biology and Philosophy*, *18*, 53–93.

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences*, *27*(3), 377–396.

Grush, R. (2006). How to, and how *not* to, bridge computational cognitive neuroscience and Husserlian phenomenology of time consciousness. *Synthese*, *153*(3), 417–450.

Gurwitsch, A. (1979). *Human encounters in the social world* (F. Kersten, Trans.). Pittsburgh: Duquesne University Press. (Original work published 1931)

Harman, G. (1978). Studying the chimpanzee's theory of mind. *Behavioral and Brain Sciences*, *1*, 576–577.

Heidegger, M. (1962). *Being and time*. New York: Harper & Row. (Original work published 1927)

Herschbach, M. (2008a). *Control-theoretic models of action understanding: Simulation, theory, or both?* Poster presented at the 34th annual meeting of the Society for Philosophy and Psychology, Philadelphia, PA.

Herschbach, M. (2008b). The concept of simulation in control-theoretic accounts of motor control and action perception. In V. Sloutsky, B. Love, & K. McRae (Eds.), *Proceedings of the 30th annual conference of the Cognitive Science Society*, (pp. 315–320). Austin, TX: Cognitive Science Society.

Hobson, R. P. (2002). *The cradle of thought*. London: Macmillan.

Hornsby, J. (2000). Personal and sub-personal: A defence of Dennett's early distinction. *Philosophical Explorations*, *3*(1), 6–24.

Hurley, S. (1998). *Consciousness in action*. Cambridge, MA: Harvard University Press.

Hurley, S. (2003a). Animal action in the space of reasons. *Mind & Language*, *18*(3), 231–256.

Hurley, S. (2003b). Making sense of animals: Interpretation vs. architecture. *Mind & Language*, *18*(3), 273–280.

Hurley, S. (2005). The shared circuits hypothesis: A unified functional architecture for control, imitation, and simulation. In S. Hurley, & N. Chater (Eds.), *Perspectives on imitation* (pp. 177–193). Cambridge, MA: MIT Press.

Hurley, S. (2006). Active perception and perceiving action: The shared circuits model. In T. S. Gendler, & J. Hawthorne (Eds.), *Perceptual experience* (pp. 205–259). New York: Oxford University Press.

Hurley, S. (2008). The shared circuits model: How control, mirroring and simulation can enable imitation, deliberation, and mindreading. *Behavioral and Brain Sciences*, *31*, 1–58.

Husserl, E. (1999). *Cartesian meditations: An introduction to phenomenology* (D. Cairns, Trans.). Dordrecht: Kluwer. (Original work published 1950)

Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.

Hutto, D. D. (2004). The limits of spectatorial folk psychology. *Mind and Language*, *19*(5), 548–573.

Hutto, D. D. (2007). The narrative practice hypothesis: Origins and applications of folk psychology. *Royal Institute of Philosophy Supplement*, *82*, 43–68.

Hutto, D. D. (2008a). *Folk psychological narratives: The sociocultural basis of understanding reasons*. Cambridge, MA: MIT Press.

Hutto, D. D. (2008b). Limited engagements and narrative extensions. *International Journal of Philosophical Studies*, *16*(3), 419–444.

Iacoboni, M., Lieberman, M. D., Knowlton, B. J., Molnar-Szakacs, I., Moritz, M., Throop, C. J., et al. (2004). Watching social interactions produces dorsomedial prefrontal and medial parietal BOLD fMRI signal increases compared to a resting baseline. *NeuroImage*, *21*(3), 1167–1173.

Jack, A. I., & Roepstorff, A. (2003). Trusting the subject? Vol. 1 [Special issue.]. *Journal of Consciousness Studies*, *10*(9–10).

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*(1), 32–38.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, *89*(1), 25–41.

Lutz, A., & Thompson, E. (2003). Neurophenomenology: Integrating subjective experience and brain dynamics in the neuroscience of consciousness. *Journal of Consciousness Studies*, *10*(9–10), 31–52.

Maibom, H. L. (2003). The mindreader and the scientist. *Mind & Language*, *18*(3), 296–315.

Maibom, H. L. (2007). Social systems. *Philosophical Psychology*, *20*(5), 557.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: W.H. Freeman.

McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(20), 11832–11835.

McDowell, J. (1994). The content of perceptual experience. *Philosophical Quarterly*, *44*(175), 190–205.

McGovern, K., & Baars, B. J. (2007). Cognitive theories of consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge handbook of consciousness* (pp. 177–205). Cambridge, England: Cambridge University Press.

Menary, R. (Ed.). (2006). *Radical enactivism: Intentionality, phenomenology and narrative; focus on the philosophy of Daniel D. Hutto*. Amsterdam: John Benjamins.

Merikle, P. M., & Daneman, M. (1998). Psychological investigations of unconscious perception. *Journal of Consciousness Studies*, *5*(1), 5–18.

Merleau-Ponty, M. (2002). *Phenomenology of perception* (C. Smith, Trans.). London: Routledge. (Original work published 1945)

Mitchell, J. P., Heatherton, T. F., & Macrae, C. N. (2002). Distinct neural systems subserve person and object knowledge. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(23), 15238–15243.

Nichols, S., & Stich, S. P. (2003). *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds*. Oxford, England: Oxford University Press.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259.

Noë, A. (2004). *Action in perception*. Cambridge, MA: MIT Press.

Noë, A., & Thompson, E. (2004). Are there neural correlates of consciousness? *Journal of Consciousness Studies*, *11*(1), 3–28.

Oberman, L. M., Pineda, J. A., & Ramachandran, V. S. (2007). The human mirror neuron system: A link between action observation and social skills. *Social Cognitive and Affective Neuroscience*, *2*(1), 62–66.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*(5719), 255–258.

Overgaard, M., Gallagher, S., & Ramsøy, T. Z. (2008). An integration of first-person methodologies in cognitive science. *Journal of Consciousness Studies*, *15*(5), 100–120.

Overgaard, S. (2005). Rethinking other minds: Wittgenstein and Levinas on expression. *Inquiry*, *48*(3), 249–274.

Penn, D. C., & Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a "theory of mind." *Philosophical Transactions of the Royal Society Series B*, *362*, 731–744.

Perner, J., & Ruffman, T. (2005). Infants' insight into the mind: How deep? *Science*, *308*(5719), 214–216.

Perry, J. (1993). *The problem of the essential indexical and other essays*. New York: Oxford University Press.

Pessoa, L., Thompson, E., & Noë, A. (1998). Finding out about filling-in: A guide to perceptual completion for visual science and the philosophy of perception. *Behavioral and Brain Sciences*, *21*(6), 723–802.

Petitot, J., Varela, F. J, Pachoud, B., & Roy, J.-M. (Eds.) (1999). *Naturalizing phenomenology: Issues in contemporary phenomenology and cognitive science*. Stanford, CA: Stanford University Press.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526.

Ramsey, W. (2007). *Representation reconsidered*. Cambridge, England: Cambridge University Press.

Ratcliffe, M. (2006a). “Folk psychology” is not folk psychology. *Phenomenology and the Cognitive Sciences*, *5*, 31–52.

Ratcliffe, M. (2006b). Phenomenology, neuroscience, and intersubjectivity. In H. L. Dreyfus, & M. A. Wrathall (Eds.), *A companion to phenomenology and existentialism* (pp. 329–345). Malden, MA: Blackwell.

Ratcliffe, M. (2007). *Rethinking commonsense psychology: A critique of folk psychology, theory of mind and simulation*. Basingstoke, Hampshire, England: Palgrave Macmillan.

Ratcliffe, M. (2008). Farewell to folk psychology: A response to Hutto. *International Journal of Philosophical Studies*, *16*(3), 445–451.

Ratcliffe, M., & Hutto, D. D. (2007). Introduction. In D. D. Hutto, & M. Ratcliffe (Eds.), *Folk psychology re-assessed* (pp. 1–22). Dordrecht, The Netherlands: Springer.

Roepstorff, A., & Jack, A. I. (2004). Trusting the subject? Vol. 2. [Special issue.]. *Journal of Consciousness Studies*, *11*(7–8).

Roy, J., Petitot, J., Pachoud, B., & Varela, F. J. (1999). Beyond the gap: An introduction to naturalizing phenomenology. In J. Petitot, F. J. Varela, B. Pachoud, & J. Roy (Eds.), *Naturalizing phenomenology: Issues in contemporary phenomenology and cognitive science* (pp. 1–80). Stanford, CA: Stanford University Press.

Ruffman, T., & Perner, J. (2005). Do infants really understand false belief? *Trends in Cognitive Sciences*, *9*(10), 462–463.

Ryle, G. (1984). *The concept of mind*. Chicago: University of Chicago Press. (Original work published 1949)

Saxe, R. (2005). Against simulation: The argument from error. *Trends in Cognitive Sciences*, *9*(4), 174–179.

Scheler, M. (1954). *The nature of sympathy* (P. Heath, Trans.). London: Routledge & Kegan Paul.

Schilbach, L., Wohlschlaeger, A. M., Kraemer, N. C., Newen, A., Shah, N. J., Fink, G. R., et al. (2006). Being with virtual others: Neural correlates of social interaction. *Neuropsychologia*, *44*(5), 718–730.

Slaughter, V., & Repacholi, B. (2003). Individual differences in theory of mind: What are we investigating? In B. Repacholi, & V. Slaughter (Eds.), *Individual*

differences in theory of mind: Implications for typical and atypical development (pp. 1–12). Hove, England: Psychology Press.

Slors, M. (2008). The importance and limits of phenomenological philosophy of mind. *Abstracta, Special Issue II*, 34–44.

Smith, D. W., & Thomasson, A. L. (Eds.) (2005). *Phenomenology and philosophy of mind*. Oxford, England: Oxford University Press.

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, *18*(7), 587–592.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition*. Oxford: Blackwell. (Original work published 1986)

Sperber, D., & Wilson, D. (2002). Pragmatics, modularity and mind-reading. *Mind & Language*, *17*(1–2), 3–23.

Spiers, H. J., & Maguire, E. A. (2006). Spontaneous mentalizing during an interactive real world task: An fMRI study. *Neuropsychologia*, *44*(10), 1674–1682.

Stein, E. (1964). *On the problem of empathy* (W. Stein, Trans.). The Hague: M. Nijhoff.

Stich, S., & Nichols, S. (1992). Folk psychology: Simulation or tacit theory? *Mind and Language*, *7*(1–2), 35–71.

Stich, S., & Nichols, S. (1995). *Second thoughts on simulation*. Cambridge: Blackwell.

Stich, S., & Ravenscroft, I. (1994). What *is* folk psychology? *Cognition*, *50*, 447–468.

Subbotsky, E. V. (1993). *Foundations of the mind: Children's understanding of reality*. Cambridge, MA: Harvard University Press.

Tager-Flusberg, H. (2005). What neurodevelopmental disorders can reveal about cognitive architecture: The example of theory of mind. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind: Structure and contents* (pp. 272–288). New York: Oxford University Press.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318.

Thompson, E. (2001). Empathy and consciousness. *Journal of Consciousness Studies*, *8*(5–7), 1–32.

Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind*. Cambridge, MA: Harvard University Press.

Thompson, E., Noë, A., & Pessoa, L. (1999). Perceptual completion: A case study in phenomenology and cognitive science. *Naturalizing phenomenology: Issues in contemporary phenomenology and cognitive science* (pp. 161–195). Stanford, CA: Stanford University Press.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Trevarthen, C., & Hubley, P. (1978). Secondary intersubjectivity: Confidence, confiding and acts of meaning in the first year. In A. Locke (Ed.), *Action, gesture, and symbols: The emergence of language* (pp. 183–229). London: Academic Press.

Varela, F. J. (1996). Neurophenomenology: A methodological remedy for the hard problem. *Journal of Consciousness Studies*, *3*(4), 330–349.

Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happé, F., Falkai, P., et al. (2001). Mind reading: Neural mechanisms of theory of mind and self-perspective. *NeuroImage*, *14*(1), 170–181.

Warneken, F., Chen, F., & Tomasello, M. (2006). Cooperative activities in young children and chimpanzees. *Child Development*, *77*(3), 640–663.

Waskan, J. A. (2006). *Models and cognition: Prediction and explanation in everyday life and in science*. Cambridge, MA: MIT Press.

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, *75*(2), 523–541.

Wheeler, M. (2005). *Reconstructing the cognitive world: The next step*. Cambridge, MA: MIT Press.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*, 103–128.

Wittgenstein, L. (1980). *Remarks on the philosophy of psychology* (C. G. Luckhardt, M. A. E. Aue, Trans.). Chicago: University of Chicago Press.

Woolley, J. D. (2006). Verbal-behavioral dissociations in development. *Child Development*, *77*(6), 1539–1553.

Yoshimi, J. (2007). Mathematizing phenomenology. *Phenomenology and the Cognitive Sciences*, *6*(3), 271–291.

Zahavi, D. (2001). Beyond empathy: Phenomenological approaches to intersubjectivity. *Journal of Consciousness Studies*, *8*(5–7), 151–167.

Zahavi, D. (2004a). The embodied self-awareness of the infant: A challenge to the theory-theory of mind? In D. Zahavi, T. Grünbaum, & J. Parnas (Eds.), *The structure and development of self-consciousness: Interdisciplinary perspectives* (pp. 35–63). Amsterdam: John Benjamins.

Zahavi, D. (2004b). Phenomenology and the project of naturalization. *Phenomenology and the Cognitive Sciences*, *3*, 331–347.

Zahavi, D. (2005). *Subjectivity and selfhood: Investigating the first-person perspective*. Cambridge, MA: MIT Press.

Zahavi, D. (2007). Expression and empathy. In D. D. Hutto, & M. Ratcliffe (Eds.), *Folk psychology re-assessed* (pp. 25–40). Dordrecht, The Netherlands: Springer.

Zahavi, D. (2008). Simulation, projection and empathy. *Consciousness and Cognition*, *17*(2), 514–522.

Zahavi, D. (2009). [Review of the book *Mind in life: Biology, phenomenology, and the sciences of mind*, by E. Thompson]. *Husserl Studies*, *25*, 159–168.

Zahavi, D. (2010). Naturalized phenomenology. In S. Gallagher, & D. Schmicking (Eds.), *Handbook of phenomenology and cognitive science* (pp. 3–20). Dordrecht: Springer.