Words About Young Minds

The Concepts of Theory, Representation, and Belief in Philosophy and Developmental Psychology

Eric Schwitzgebel

B.A. (Stanford University) 1990

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

Doctor of Philosophy
in

Philosophy

in the graduate division of the
University of California at Berkeley

Committee in charge:

Professor Elisabeth A. Lloyd (chair)
Professor John R. Searle
Professor Alison Gopnik (psychology)

May, 1997

The dissertation of Eric Schwitzgebel is approved.

_____

Professor Elisabeth A. Lloyd (chair) Date

_____

Professor John R. Searle Date

_____

Professor Alison Gopnik Date

University of California at Berkeley
May 1997

Words About Young Minds:
The Concepts of Theory, Representation, and Belief
in Philosophy and Developmental Psychology

May 1997

Abstract

Words About Young Minds:
The Concepts of Theory, Representation, and Belief
in Philosophy and Developmental Psychology

Eric Schwitzgebel

Doctor of Philosophy in Philosophy
University of California at Berkeley
Professor Elisabeth A. Lloyd, Chair

I examine three philosophically important concepts that play a foundational role in developmental psychology: theory, representation, and belief. I describe different ways in which the concepts have been understood and present reasons why a developmental psychologist, or a philosopher attuned to cognitive development, should prefer one understanding of these concepts over another.

I take up the concept of theories with an eye to recent debate in psychology over whether the cognitive development of young children can fruitfully be characterized as involving theory change. I propose, instead, a novel account of theories intended to capture what scientific theories and everyday theories have in common. I connect theories with the emergence and resolution of explanation-seeking curiosity, and I argue that if developmental psychologists want convincingly to defend the view that young children have theories, they must look for the patterns of affect and arousal associated with such curiosity.

I begin my discussion of the concept of representation by distinguishing between two very different conceptions of

representation at work in the philosophical literature. I argue that both philosophers and psychologists have tended to conflate these two conceptions, and I examine the serious consequences of this conflation for the developmental literature on the child's understanding of mind. I suggest some empirical research that looks promising once this conceptual tangle is straightened out.

Finally, I examine the concept of belief. I provide detailed objections to Donald Davidson's claim that creatures without language, including human infants, cannot have beliefs, and I argue that the interests of both philosophers of mind and developmental psychologists are best served by a dispositional account of belief, appealing not merely to dispositions to behave, but also to dispositions to have certain kinds of subjective experiences. This account offers a satisfying resolution to several problems in philosophy and developmental psychology, including those raised by Putnam's Twin Earth case, Kripke's puzzle about belief, the phenomenon of self-deception, and conflicting data from child psychology on the development of the object concept and the child's understanding of false beliefs.

Contents

Acknowledgements

I take great pleasure in discussing philosophy in coffee houses with friends, and in trading philosophical thoughts by email. My ideas are often forged in dialogue to be refined on the page, and consequently I owe large philosophical debts to a number of people.

For their help in treating the topic of chapter two, I would like to thank fellow graduate students Carl Anderson, Herman Capellen, Andy Carpenter, and Mark Wrathall; professors Fred Dretske, P. J. Ivanhoe, Ernie Lepore, Tori McGeer, and Karsten Stueber; my psychologist buddy Penny Vinden; and my former landlady Nancy Coolidge and a shuttle driver in Long Island for their patient discussions of the ordinary usage of the word 'belief,' which I found to deviate to a surprising degree from common philosophical usage. John Heil's careful reading of and detailed comments on the chapter were immensely helpful in keeping my treatment of Davidson accurate and honest.

My thought on the topic of chapter three has been greatly improved by conversations with, among others, graduate students Eddie Cushman, Josh Dever, Mathias Frisch, and Bojana Mladenovic, and professor Martin Jones. I also received helpful comments on a shortened version of the chapter presented at the 1996 meeting of the Philosophy of Science Association.

Chapter four has been immensely improved by discussion with nearly every figure who plays an important role in it -- not only my advisors, but also Judy DeLoache, Fred Dretske, Josef Perner, and Dennis Stampe. Lauren Silver helped me get up to speed with the literature on artistic development, and Brian Foley did the same with the educational literature on the theory theory. Others who have commented usefully on the ideas in chapter three include Maggie Friend, Angel Lillard, Bojana Mladenovic, Pauline Price, Andrea Rosati, Penny Vinden, and June Wai.

The chapters on belief have received much of my most recent attention and owe quite a bit to others. I would like especially to thank John Heil and Tori McGeer in this regard, for it was in discussions with them that I first became convinced that a dispositional account of belief could work, and my thoughts on this matter owe much to their own. The ideas in this chapter have been presented at meetings of the philosophy of mind and language group at U.C. Berkeley, the Berkeley philosophy department colloquium, and as a job-talk at U.C. Riverside, and all these audiences have offered helpful criticisms and insights. My fellow graduate students Max Deutsch, Josh Dever, and John Madsen have been particularly helpful in pressing their criticisms. Others with whom I have had especially rewarding discussions include John Campbell, Martin Jones, and Ariela Lazar.

A number of people have contributed so much to the dissertation as a whole that I would like to acknowledge them separately. First, I must thank Kim Kempton, Maria Merritt, and

Leo van Munching, who, with me as the fourth, formed a dissertation discussion group that has met every week or two for the last two academic years. Not only their comments on drafts of sections, but especially their sympathetic support and comraderie has been invaluable in keeping my dissertation on course. I will greatly miss our amiable and thoughtful discussions. A less formal, more social circle of friends from the Institute of Human Development has likewise provided much moral support, especially Maggie Friend, Tori McGeer, Betty Repacholi, and Penny Vinden.

I would also like to thank John Searle, who, as the second member of my committee, gave me penetrating and insightful comments on each of my chapters in their raw, original forms. After my spending too much time reading psychology journals or philosophy of science, John served very well to force my attention back to the issues in philosophy of mind that form much of the root structure of the work in this dissertation. I must also thank him for his sympathy and patience with a treatment of belief that diverges substantially from his own.

I must also thank my good friend Tori McGeer for immensely productive philosophical discussions on all the topics of the dissertation, and many other topics besides. Many philosophical sentiments and opinions we have shared with almost preternatural exactness, so that our conversations have often taken the form of co-operative building, rather than the thrust and parry of argument and counterargument nearly universal in philosophical discussion. I leave our discussions impassioned by new ideas and

eager to write. Tori has also been for me a model of subtlety and elegance in thought and writing, and of honesty, open-mindedness, and undefensiveness in philosophical discussion.

Alison Gopnik, psychologist *cum* philosopher, drew me as much to developmental psychology as the topic itself did, with her acute nose for connections between issues developmental and philosophical and with her refreshingly pragmatic approach to philosophy. She has been uncommonly kind in opening her lab to me, giving me a foothold in the Institute of Human Development, and allowing me to observe experiments and participate in lab meetings. Her seminars and meetings of the Meno Society, organized by her, have been immensely influential in cultivating my taste for the theoretical, conceptual, and philosophical in developmental psychology. I am not the only philosopher to think that better philosophy goes on among developmental psychologists gathered by Alison than often goes on in philosophy departments. Some of the other developmental psychologists with a theoretical, philosophical bent who have been welcoming of me include professors John Flavell, Angel Lillard, Dan Slobin, and John Watson, postdocs Penny Vinden and Maggie Friend, and graduate students both in and out of Alison’s lab.

Lisa Lloyd has been a nearly perfect advisor, warm and welcoming, insightful, supportive, a careful reader with a gentle touch in criticism, astute, knowledgeable, broad-visioned, and a thoughtful, patient discussant. We often have spent four hours on a Saturday or Sunday discussing my work in detail, as well as

all manner of other philosophical issues. In discussion, there seems to be a natural harmony between us. Lisa continued to treat her work as an advisor as major priority through times of serious illness, and for this I am especially thankful. Had I been more aware of the gravity of her illness, and the importance of her own research in discovering a cure, I would certainly have made fewer demands on her time.

I must also thank my parents, Colleen Ryan and Kirk Gable, for their consistent encouragement of my work in philosophy, their willingness to support me financially when necessary, and their enthusiasm for reading my dry, philosophical prose. Finally, my greatest appreciation goes to my partner, Pauline Price, whose patience with my solitary retreats, my distraction, and my moodiness has helped keep me balanced and stable. She knows what is important.

Chapter One

Introduction to the Dissertation: Philosophy, Developmental Psychology, and Intuition

The history of philosophy is thoroughly entangled with developmental psychology. In Plato's *Meno*, Socrates applies his famous "doctrine of recollection," according to which all learning is just recollection of things antecedently known from past lives, to questions about the nature of morality and to skeptical concerns about the possibility of learning. John Locke devotes the entire first book, and much of the second book, of his *Essay Concerning Human Understanding* to an extended discussion of the origin of ideas, interweaving developmental claims about the origins of various types of ideas with philosophical claims about their nature. His discussions of the nature of words and of the origins, extent, and reality of knowledge likewise hang upon developmental theses. Philosophers reacting to Locke, such as Berkeley, Hume, and Kant, have likewise seen connections between developmental issues and various issues at the center of philosophy.

Contemporary philosophers continue to make connections with developmental psychology. Willard Quine's *Word and Object* concerns itself centrally with the learning of language from scratch (as a child or jungle anthropologist would), and Quine draws substantial philosophical conclusions about the nature of

language (e.g., about its indeterminacy) from these observations. Donald Davidson (1984) has drawn even broader conclusions about language, knowledge, and the mind from a similar starting point, and some of his views and their developmental connections will be discussed in the second chapter. Jerry Fodor (1983) has helped to revive an innatist view of the mind in philosophy, again connecting issues about the development of the mind with issues about its nature. In fact, it is hard to find a philosopher of mind, language, or epistemology who isn't committed to some view or other about the development of children.

Developmental psychology, likewise, often builds upon philosophical foundations. Like many other sciences, developmental psychology had its origins in philosophy, and much work in developmental psychology still explicitly positions itself with respect to developmental claims made by philosophers such as Locke and Kant. The work of contemporary philosophers has also had a great impact on developmental psychology. The contemporary developmental literature on the child's "theory of mind," for example, grew out of observations by philosophers on the importance of a creature's understanding false belief for its understanding of the mind (Bennett 1978; Dennett 1978; Harman 1978; Wimmer and Perner 1983), and much of the work in theory of mind still draws upon the observations of philosophers of mind such as John Searle, Fred Dretske, and Jerry Fodor. Work on language development (e.g., Markman 1989) has set itself puzzles drawn from Quine's (1960) work described above. Work on conceptual change in childhood (e.g., by Carey 1985; Gopnik and

Meltzoff 1997) has drawn upon work on conceptual change in philosophy of science (especially Kuhn 1960/1970). Work on the nature of the child's concepts (e.g., Keil 1991; Gelman and Coley 1991) has drawn on philosophical discussions of the nature of human concepts in general (e.g., Wittgenstein 1958; Putnam 1975b; Millikan 1997).

While many philosophers find themselves committed to developmental positions, or to positions that developmental psychologists have thought to have consequences for their work, few contemporary philosophers have explored the empirical side of developmental psychology in any extended way. In this dissertation, which treats philosophical issues that arise in the context of developmental psychology, I hope to help remedy this deficit. In particular, I will examine the concepts of *theory, representation,* and *belief* as they arise in recent philosophical and developmental work. These concepts play a crucial role in both disciplines.

The concept of *theory* plays a crucial and obvious role in the philosophy of science: Most philosophers of science suppose one of the primary enterprises of science, if not *the* primary enterprise, to be the construction and evaluation of theories. It is therefore almost impossible to do work in philosophy of science without discussing, in one way or another, scientific theories. For developmental psychology as well, the concept of a theory has played an important role. At least since the time of Jean Piaget, some developmental psychologists have likened the

cognitive development of children to the processes of theory change in science. Children are thought of as constructing, testing, modifying, and rejecting theories much as scientists do, and a major task of developmental psychology, on this view, is the elaboration of children's theories and the mechanisms of their change. There is currently substantial debate over the value of such a "theory theory" within developmental psychology. What is at stake is nothing less than a general picture of the nature of cognitive development.

The concept of *representation* has also played an important role in the work of philosophers and developmental psychologists. Many philosophers of mind, such as Fred Dretske (1988), John Searle (1983), and Jerry Fodor (1975, 1990) have taken representations to be among the most important components of the mind, although they have defined the term 'representation' in rather different ways, as we will see in chapter four. A number of developmental psychologists have followed them in this, and some, such as Josef Perner (1991), Alison Gopnik (Gopnik and Astington 1988), and Henry Wellman (1990), have argued that coming to understand the representational nature of mind is a major accomplishment in the preschooler's development of an understanding of minds. However, unless we have a clear understanding of what a "representation" is supposed to be, then we cannot clearly understand either the philosopher's claims about the nature of mind or the developmentalist's claims about the child's development of an understanding of mind.

Finally, the concept of *belief* also plays a crucial role in much work in both philosophy and developmental psychology. Most philosophers interested in representation regard beliefs as central cases of representations, and many begin their discussions of representation in general with discussions of belief in particular. Our ordinary, "folk" psychology is, in the view of many philosophers (e.g., Fodor 1987; Stich 1983; Searle 1992), a psychology in which belief plays a central role, and philosophical explanations of behavior that are friendly to folk psychology often appeal primarily to beliefs and desires. Developmental psychologists sympathetic to such views have regarded development of the child's understanding of belief as crucial in the development of a theory of mind in general (Perner 1991; Wellman 1990; Astington 1993). More broadly, many developmental psychologists have seen the determination of what it is a child believes about any particular subject as crucial in characterizing the child's understanding of that subject. Cognitive development is often understood, by such psychologists, as largely consisting in changes in the child's *beliefs*.

Outline of the Dissertation

The second chapter of the dissertation initiates my discussion of the concept of *belief*. In particular, I explore the question of whether infants and non-human animals, creatures without language, can have beliefs. I examine two well-known arguments against infant and animal belief advanced by Donald

Davidson, and I show how those arguments fail to establish their conclusion. I then offer a plausibility argument in favor of thinking that infants and animals have beliefs, and I describe some practical considerations suggesting that the term 'belief,' if it is to be of broad use in academia, ought not to apply exclusively to the cognition of adult human beings. A general account of belief is not offered until later in the dissertation.

The third chapter of the dissertation treats the concept of a *theory*. In particular, this chapter is concerned with the debate within developmental psychology over how much the cognitive development of children is like theory change in science. Useful debate on this topic requires a clear understanding of what it would be for a child to have a theory. I argue that existing accounts of theories within philosophy of science and developmental psychology either are less precise than is ideal for the task or cannot capture everyday theorizing of the sort that children, if they theorize, must do. I then propose an account of theories that ties theories and explanation very closely together, treating theories primarily as products of a drive to explain. I clarify some of the positions people have taken regarding the "theory theory" of development, and I conclude by proposing that psychologists interested in the theory theory look for patterns of affect and arousal in development that would accompany the existence of a drive to explain.

I begin chapter four by distinguishing two very different conceptions of *representation* at work in the philosophical

literature. On the first, “contentive” conception (found, for example, in John Searle and Jerry Fodor), something is a representation, roughly, just in case it has “propositional content”; on the second, “indicative” conception (found, for example, in Fred Dretske), representations must not only have content but must also have the function of indicating something about the world. I argue that the philosopher Dennis Stampe conflates these two conceptions in a seminal paper of his on representation, and that Alison Gopnik and Josef Perner conflate these conceptions in their discussions of the child’s understanding of the mind. The latter conflation, I argue, leads Gopnik and Perner to think that when the child comes to appreciate the nature of misrepresentation at age four, the child must also undergo some change in her understanding of desire. This chapter, like the previous one, concludes with some suggestions for empirical research. In particular, I argue that it is an open question whether the child understands indicative representation generally at age four, and that one useful test of this hypothesis would look at the child’s understanding of representational art.

Chapter five returns to the topic of belief. In this chapter, I describe some of the desiderata of an account of belief, and I argue for the existence of “in-between” states of believing, in which a subject cannot accurately be described either as fully believing or fully failing to believe the proposition in question. I also describe in some detail the container metaphor for belief, quite popular now in philosophy of

mind, and I suggest that some of the images evoked by this metaphor may draw us toward a view of belief that, on reflection, we would not want to accept.

In chapter six I offer a "phenomenal, dispositional" account of belief. The account is dispositional because it treats believing as matching to an appropriate degree a stereotypical set of dispositions. The account is phenomenal, because unlike dispositional accounts as typically conceived, the dispositions belonging to that stereotypical set include dispositions to have certain sorts of phenomenal experiences. One of the primary virtues I claim for this account is its facility in handling cases of in-between believing, and I describe its application to a number of cases of in-between believing. The last two sections of the chapter are intended to forestall possible objections to the account. In the first of those sections, I defend the view that appeal to the causes of a belief is not necessary for full characterization of that belief. In the last section, I argue that beliefs conceived dispositionally can both cause and explain phenomenology and behavior.

In chapter seven, I apply the account of belief just developed to two puzzle cases in philosophy and two puzzle cases in developmental psychology. I argue that both Saul Kripke's "Puzzle about Belief" and the self-deception literature in philosophy suffer from a failure to recognize the legitimacy of describing a subject as being in an in-between state of believing. With my dispositional account of belief in hand, the cases described by Kripke and by philosophers interested in self-

deception no longer look so puzzling. I then argue that the developmental literature on the child's understanding of object permanence, as well as a paper by Wendy Clements and Josef Perner, similarly suffer from a failure to recognize the legitimacy of describing the child as in an in-between state of believing regarding the topics at hand. One ought, in fact, to *expect* that the gradual development of new competencies and new understandings of the world will move children gradually through periods in which they cannot accurately be described as either fully believing or fully failing to believe the propositions expressing the knowledge they unequivocally have at the end of the process.

Chapter eight briefly reviews the dissertation, with a particular eye to the practical benefits of my work for the fields of philosophy and developmental psychology.

The Role of Analysis and Intuition in This Dissertation

For two of the concepts discussed in this dissertation, *theory* and *belief,* I provide a novel analysis, and for one of the concepts, *representation,* I provide a clarification of some differences between existing analyses. In the course of doing this conceptual work, a number of practical decisions must be made that reflect my view of the aims of conceptual analysis. In the last few decades, most philosophers have been too quiet about the values guiding their conceptual analyses, but in mid-century, a number of philosophers quite explicitly debated what these aims

should be. So, for example, Norman Malcolm (1942) argued that philosophical analyses of words and concepts must cleave precisely to ordinary language usages, and that only confusion and falsity is to be gained by any attempt at conceptual modification and linguistic redefinition (at least by philosophers). Less extreme "ordinary language" philosophers such as John Austin (1956) simply recommended close study of and adherence to ordinary language as a fruitful, guiding technique for philosophers. In opposition, a number of people working in philosophy of science, such as Rudolf Carnap (1962) and Carl Hempel and Paul Oppenheim (1948), saw "explication" as a central project of philosophy. Explication was defined as the process of transforming an inexact concept from ordinary language (the *explicandum*) into a more exact concept for philosophical and scientific use (the *explicatum*). Carnap (1966, p. 5) describes four goals that must be balanced in explication: (1) similarity of the explicatum to the explicandum, (2) exactness, (3) fruitfulness, and (4) simplicity.

The approach taken in this dissertation has more in common with Carnap's approach than with Malcolm's. My aim is to assist philosophers and developmental psychologists in developing concepts that will be *pragmatically useful* for their academic research. While it is definitely desirable to treat concepts in a way that matches up to some extent with pre-theoretical, ordinary concepts -- for ease of understanding, if nothing else -- assuring such a match cannot be the final value of

pragmatically-oriented conceptual work. Simplicity, fruitfulness, coherence with important distinctions and divisions in the field, and precision (and sometimes vagueness in the right places -- see my chapters on belief) must all be considered as goals in concept-tinkering, and people with different interests may reach different conclusions about how these factors are all to be balanced and thus about how best to analyze a particular concept. Ordinary-language analysis is more like analysis in the strict sense of breaking apart and displaying what is already present. My project is not so much to describe existing use as to make recommendations for future use; something new is constructed and offered up to take the place of, or to give definiteness to, an existing vague or muddled concept.

Taking this pragmatic approach to conceptual analysis requires a certain willingness to say things that run contrary to our pre-theoretical intuitions. At the very least, we should be unsurprised if an explication strains some of our linguistic intuitions about when it is and is not appropriate to use a certain word -- a natural consequence of the effort to *adjust* our understanding of particular concepts and the words attached to them. Some of the claims in this dissertation may diverge from the reader's intuition in other respects as well, when those intuitions conflict with conclusions established on the basis of empirical evidence or philosophical inquiry.

Since the charge that a philosophical claim is "counter-intuitive" is often employed as though being counter-intuitive by

itself were reason enough to abandon a claim, I would like briefly to discuss the role I see intuition playing in this dissertation. Philosophers all too often are insufficiently deliberative about the assumptions involved in condemning a position for having counter-intuitive consequences, and I want to give at least momentary pause to the reader who might be inclined to leap in immediately with such assessments of the work to come.[1]

The conclusion one is usually meant to draw from the charge that a claim is counter-intuitive is that that claim must be (or probably is) false. The typical role of a charge of counter-intuitiveness, accordingly, is as the penultimate step in a *reductio.* But certainly this form of argument will work only in domains for which intuition is taken to be a reliable guide. No one argues, any more, that it is counter-intuitive to claim that the Earth revolves around the Sun, and therefore there must be something wrong with our celestial mechanics. Nor does anyone argue that it is counter-intuitive to assert that things gain mass as they approach the speed of light, and therefore Einstein's theory of relativity stands in need of correction. But in philosophy the accusation of counter-intuitivity is taken seriously. What is supposed to be the difference?

In certain fields, intuitions are the foundations atop which all theories must be built. In linguistics, intuitions about grammaticality are an important part of the raw data for theories of grammar; if a grammatical theory produces predictions that too

[1] Elements of this discussion will also appear in Gopnik and Schwitzgebel (1997).

seriously violate our grammatical intuitions, we must reject the theory. (Nevertheless, things are not entirely simple, as is made evident by such famous sentences as "The horse raced past the barn fell"[2].) One might argue that the same is true in moral philosophy: We have certain intuitions about what is moral and what is not moral, and it is the business of moral philosophy to construct theories that account for the accuracy of these intuitions and organize them into a workable structure. Still, it is controversial whether this is how moral philosophy does (or should) work.

Intuitions are an important part of the data in philosophy of mind as well. We make intuitive judgments about our minds, about our experiences, perceptions, and internal states. However, it should be noticed that the data that must be accommodated in philosophy of mind are the intuitive *judgments* that some such propositions P, Q, etc. are true, which leaves as an open question whether P, Q, etc. are *actually* true. In this regard, intuitions play a slightly different role in philosophy of mind than in linguistics or moral philosophy as conceived above: In the latter fields, when people have an intuition that X is F (e.g., X is ungrammatical or immoral), the datum to be accounted for is the F-ness of X, and the occurrence of the intuition itself is only attended to in a secondary way if at all, while in philosophy of mind the reverse is true. That we *make* certain intuitive judgments is an undeniable fact that philosophers of

[2] That this sentence is grammatical can be seen by comparing it to the similarly-structured sentence "The man hit with the rock shouted."

mind must accommodate; whether those intuitive judgments are *right* is a separate question altogether.

Intuitions are reliable indicators of the adequacy of grammatical, moral, and philosophical theories of the sort described above in a *constitutive* way: They are the very material that the theories seek to organize. Intuitions may be reliable in a non-constitutive way also. The expert chess-player may have an intuition that one chess move is better than another, even if she cannot articulate exactly why. If I develop a theory of chess that is meant to classify certain types of moves as good and others as poor, and the theory runs contrary in a range of cases to Gary Kasparov's intuitions, I have good reason to be concerned. His intuitions have been honed by long practice and have been employed in brilliant chess play. On the other hand, it is not *impossible,* the way it is in the grammatical case, that my theory is right and Kasparov is broadly mistaken (for example, if everyone else is more terribly mistaken). A theory of chess that violates Kasparov's intuitions may be unlikely to succeed, but in simply in violating those intuitions it does not conflict with a piece of the data the theory is attempting to organize. The theory is not about Kasparov's intuitions; it is about chess.

We are all experts, in a pragmatic sort of way, in everyday psychology, and perhaps for this reason, counter-intuitive claims in psychology or philosophy of mind should be regarded as *prima facie* less plausible than intuitive claims, just as we would regard as *prima facie* less plausible a theory of chess that ran

counter to a grandmaster's intuitions. After all, our psychological intuitions are grounded in wide experience of our own minds and the minds of others. Nonetheless, it does not follow that our psychological intuitions are infallible, just as a grandmaster's intuitions about chess aren't infallible. Nor need our intuitions even be entirely coherent. (At the end of section one in chapter four, and in chapters five and six, I will point out some places in which our intuitions may pull in conflicting directions.) As psychological science has matured, we have become more confident in leaving intuitions behind when they conflict with well-supported psychological claims, as for example in the cases of blindsight (Weiskrantz 1986) and attribution error (Nisbett and Ross 1980). So, although we may justifiably take reflective psychological intuition as a good preliminary guide, we no longer take it as the final authority about the mind.

One might argue that all judgments rest, ultimately, in intuition, and that therefore there can really be no court of appeal beyond that of intuition. Even, however, if we accept the premise that all judgments do ultimately rest on intuition (however such a claim is to be spelled out), the conclusion either does not follow or is irrelevant to the issue at hand. We don't take seriously our intuitions about the details of physics, and we are right not to. Therefore, either intuition is not always the final court of appeal, or it is, but certain counter-intuitive judgments are nonetheless acceptable in the face of strong evidence (which might be thought of, on this view, as

stronger intuitions that conflict with it), and thus the argument cannot without further work establish that we must adhere to any specific intuitions we might have in discussing the mind in particular.

All this said, I do not think that anything I will defend in this dissertation rebels too violently against our intuitions. In my arguments and my analyses, I will be guided by what I hope to be a well-considered balance of reflective intuition, philosophical and psychological theory, empirical data, and a pragmatic aesthetic of simplicity.

Chapter Two

A Defense of the View that Infants and Animals Have Beliefs

We normally treat infants and non-human animals as though they have beliefs and desires. We predict and explain their actions on the basis of what we think they want and what they know about how to get the things they want. We think of them as sometimes disappointed, surprised, afraid, and so forth, as a result of their hopes and expectations about the world. We describe them with character traits that seem to presuppose their possession of beliefs and desires -- as sneaky, clever, or ill-tempered, for example. A number of developmental psychologists and cognitive ethologists have allowed such belief-desire terminology to come into their scientific work. For those with a philosophical turn of mind the question naturally arises, is it true to say of such creatures that they have this range of cognitive states, or is it merely a convenient (but perhaps misleading) way of talking?

In this chapter I will defend the view that we are not merely speaking loosely or metaphorically when we attribute beliefs to infants and animals. (I think a similar argument can be made with respect to desires and the other so-called “propositional attitudes,” but I shall focus my attention solely upon belief.) Developmental psychologists and those who study some of the more

cognitively sophisticated mammals such as dogs and apes, should feel no compunction, I think, in using these terms from folk psychology to describe the cognitive lives of the creatures they study. Babies and Saint Bernards have beliefs.

Not all philosophers share my view on the matter, of course. Descartes held that animals had no souls and hence no beliefs (1637/1980). Paul Churchland (1981) argues that *nobody* has beliefs, and so, of course, infants and animals don't. I will not discuss Descartes' or Churchland's arguments in any detail. Both require the acceptance of larger pictures that I will simply suppose the reader to reject. Descartes' position depends upon a particular dualist view of the soul and the mind. Churchland's position depends upon his rejection of "folk psychology." If the reader is attached to either of these views, what I say in this chapter will no doubt seem beside the point.

I take my primary opponent on the subject of infant and animal belief to be Donald Davidson. I focus on infant and animal *belief* here because Davidson does -- but, like Davidson, I think belief and desire must come as a pair. It would hardly make sense to preserve one half of this duo while rejecting the other.

Davidson has two arguments against infant and animal belief, both of which appeared originally in "Thought and Talk" (1975/1984) and were later refined in "Rational Animals" (1982b). I devote one section each to rebutting these arguments and a third section to providing my own positive argument on behalf of

infant and animal belief. I devote so much attention to Davidson for two reasons. First, Davidson's papers are probably the most influential contemporary attacks on infant and animal belief, so it is worth examining them to see what attraction they hold. And second, it is my hope that once Davidson's arguments are shown to be faulty, the reader will naturally be drawn to the view I defend, and a large part of the work will already be done for me. Nothing will remain to stand in the way of our natural inclination to take seriously the attribution of beliefs, desires, and all the usual organs of folk psychology to infants and animals.

Before heading into the main body of this chapter, I would like to give the reader a rough sense of how I see the debate over whether infants and animals have beliefs. In my view, the question has two components which are sometimes not clearly distinguished. First, what are the conditions under which a creature may truly be said to have beliefs? Second, do real, living gorillas and six-month-olds satisfy these conditions? Davidson's attention is properly (for a philosopher) focussed on the first of these two questions, as mine will be, although the second question cannot go completely without notice. Davidson's hope, and mine, is that given our respective answers to the first question, the answer to the second will be obvious and require no subtle empirical research.

But what kind of question is the first question, the question about the conditions under which a creature may be said to have beliefs? Perhaps this question will strike some philosophers as

a request for the conceptual analysis of a piece of ordinary language, the word ‘belief,’ to be answered with a set of necessary and sufficient conditions which capture our ordinary intuitions about the extension of the term. I do not see the matter this way.

To begin with, the word ‘belief’ as it has been used by philosophers of mind and cognitive psychologists is a technical term, and its usage may even be somewhat at variance with ordinary usage (although many philosophers would deny it). I have observed, for example, that people seem to be reluctant to use the word ‘believe’ or ‘belief’ except in contexts of discussing deeply held, controversial convictions, such as religious or political convictions, and in contexts of uncertainty or disagreement. Possibly ‘I believe’ is also used simply to indicate deference (as when the ticket taker says “I believe your seat in in the third row, sir”). The verb ‘think’ in ordinary English may come closer to the philosopher’s sense of ‘believe,’ but there is no good nominal counterpart, since the word ‘thought’ has a rather different sense from the philosopher’s ‘belief.’[1]

Facts about ordinary usage aside, there seems to me no good reason *not* to treat the word ‘belief’ as a technical term for philosophers of mind and cognitive psychologists and thus give it whatever meaning and use best suits our purposes as practioners of these disciplines. Of course, if the meaning we give it is

too much at variance with previous meanings, people are apt to be confused by our use of the term, so there is a good practical reason not to stray too far from what others have said. But, as with any decision about the use of a technical term, the decision about the use of the word 'belief' is a *practical decision*, guided by *practical considerations*.

It is in this light that I wish to view the question of what the conditions are under which it would be appropriate to say that a creature has beliefs. It is my position that for most of the purposes to which philosophers of mind and cognitive psychologists may wish to employ the word 'belief,' it makes sense to regard infants and animals as having beliefs. This is a strong claim: Not only do I think that infants and animals really do have beliefs in the sense of 'belief' I endorse (and will defend in Chapter Six), but I also think that any attempt to redefine the term 'belief' so as to escape this conclusion is apt to fail as a general-purpose definition of the term.

[1] Nelson (1983) also argues that ordinary usage of the word "belief" implies a kind of "two-mindedness" about matters -- an implication absent from most philosophers' accounts of belief and its relation to action.

1. Faults in Davidson's First Argument Against Belief Without Language

Davidson claims that infants and animals, lacking language, cannot have beliefs. He defends this view primarily in two articles, "Thought and Talk" (1975/1984) and "Rational Animals" (1982b). The two papers are similar in structure. Both offer a preliminary argument and then proceed to a shorter main argument. Both the preliminary and the main arguments remain essentially the same between the two articles, although the later article contains a few twists not present in the earlier paper. In this section I will examine and criticize Davidson's first, preliminary argument as it appears in the two papers.

Both of Davidson's arguments work on the presupposition that infants and animals are incapable of language. Some have attacked Davidson on just this point. Vicki Hearne (1982), for example, has argued that well-trained dogs and horses do have language. I say "fetch!" and the dog fetches. I say "stay!" and the dog stays. The dog and I communicate with each other by means of verbal commands on my part and actions and postures on both our parts. Even more has been claimed for signing apes, such as Washoe and Koko, who seem to be capable of producing and understanding a couple hundred simplified signs from American Sign Language and who may even be able to put them together in novel, meaningful ways.[2]

I will not pursue this particular line of attack against Davidson. First of all, I am not sure it can easily be adapted

[2] Savage-Rumbaugh (1986) provides a good discussion of this topic.

to apply to very young infants, whose communicative capacity seems to be somewhat less than that of a signing ape or a well-trained dog, but who, nonetheless, I want to say have beliefs. Additionally, there seems to be a perfectly good sense of 'language' on which it is fair to say that dogs and infants before they produce their first words do not have language, and on which one may even be able to raise doubts about the signing apes. In any case, I am willing to grant Davidson the point. My interest is not in debating over what ought to count as an instance of language use.

In both "Thought and Talk" and "Rational Animals," Davidson begins his argument with a retelling of Norman Malcolm's (1973) story about a certain dog -- I will call him "Ajax," after my neighbor's dog. The story is intended by Malcolm to show that dogs "think."[3] Here is the story.

> Suppose our dog is chasing the neighbor's cat. The latter runs full tilt toward the oak tree, but suddenly swerves at the last moment and disappears up a nearby maple. The dog doesn't see this maneuver and on arriving at the oak tree he rears up on his hind feet, paws at the trunk as if trying to scale it, and barks excitedly into the branches above. We who observe this whole episode from a window say, "He thinks that the cat went up that oak tree" (1973, p. 13).

Malcolm seems to be happy with an ordinary language argument for the view that dogs think, but Davidson is willing to consider the possibility that ordinary language leads us astray in this case. Davidson's argument begins with the observation that, presumably,

[3] Although Davidson represents Malcolm as intending to use the story to show that dogs have beliefs, Malcolm is actually quite careful to phrase his claim as a claim that dogs "think," which he distinguishes from "having thoughts." The latter, Malcolm argues, is not possible without language. It is not clear from this story what Malcolm would say

if Ajax has a belief, it must be a belief with some specific content or other. The question arises, then, what precisely this content is. Consider a variety of expressions that might be taken to refer to the oak tree in question, such as ‘the oldest tree in sight,’ ‘the only tree eight meters from the house,’ or ‘the tree planted by Aunt Janet.’ Davidson assumes, and I think it is plausible to assume, that the belief that the cat ran up the oldest tree in sight is not the same as the belief that the cat ran up the tree planted by Aunt Janet. A person could easily believe one without believing the other. In general, it seems plausible to suppose that two sentences may describe different beliefs even if those sentences differ only in having different ways of picking out the same referents.

It is important to make this point carefully. Consider the following sentences:

(1.) The cat went up the oldest tree in sight.

(2.) Mary believes the cat went up the oldest tree in sight.

(3.) The cat went up the tree planted by Aunt Janet.

(4.) Mary believes the cat went up the tree planted by Aunt Janet.

The truth value of the first sentence cannot be changed by substituting for ‘the oldest tree in sight’ a term that picks out the same referent as that term -- in our example, ‘the tree planted by Aunt Janet.’ Given that ‘the oldest tree in sight’

---

about beliefs. As far as I can tell, Davidson uses “think” and “believe” more or less interchangeably.

refers to the same tree as ‘the tree planted by Aunt Janet,’ sentences (1.) and (3.) must have the same truth value. Sentences such as these, in which the substitution of co-referring terms is truth-preserving, are usually called *referentially transparent*.

Sentences (2.) and (4.), on the other hand, are *referentially opaque*. Truth value is not always preserved under substitution of co-referring expressions. Even if it is the case that ‘the oldest tree in sight’ picks out the same tree as ‘the tree planted by Aunt Janet,’ sentence (2.) may be true while sentence (4.) is false, or vice versa -- if, for example, Mary does not know that the tree in question was planted by Aunt Janet.

This fact about belief ascriptions, of course, mirrors a fact about the beliefs being ascribed. Beliefs seem to have very specific contents: Mary’s belief is definitely that the cat went up the oldest tree in sight, not that the cat went up the tree planted by Aunt Janet. Searle (1992) calls this feature of beliefs *aspectual shape*.

If we accept (as I think we should) that belief attribution sentences exhibit referential opacity and that beliefs themselves have aspectual shape, it begins to look like a tricky matter to determine exactly what it is our dog Ajax believes. Certainly it seems a mistake to ascribe to him the belief that the cat went up the oldest tree in sight, since it is doubtful that dogs do much in the way of assessing tree age. Is it even right to say that he believes the cat went up the *tree?* What do dogs know about

trees? Davidson holds that in order to have the belief that the cat went up the tree, a dog (or any creature) must be able to believe of objects *that* they are trees -- and this latter kind of belief requires that dogs know all kinds of things about trees. Examples Davidson gives include: that they are growing things, that they need soil and water, that they have leaves or needles, that they burn (1982b, p. 320). This idea that one belief is not possible without a network of other beliefs to give the first belief content Davidson sometimes calls "holism."

Davidson's argument, then, is essentially the following.[4] If we wish intelligibly to ascribe a belief to a dog, we must decide first exactly *what* belief to ascribe. But to determine exactly what belief is appropriate to ascribe to a dog, we must make judgments about a wide range of other beliefs the dog might be taken to have. Soon we will find ourselves in dubious territory, forced to make decisions about whether, for example, Ajax believes that trees need soil to grow -- decisions it seems we could have no rational basis to make. Without a language, Davidson thinks, a creature's behavior cannot have the kind of richness and diversity necessary to support the required judgments. There's just no way to pick out, and quite probably no real fact of the matter, which among a set of sentences with co-referential terms are the sentences that may accurately be said to capture the creature's beliefs. Something is amiss,

[4] Heil (1992) gives a clear and helpful exposition of it.

then, in the project of trying to ascribe beliefs to such creatures in the first place.[5]

I have several criticisms of Davidson's first argument as presented here. First, it is not clear exactly what its conclusion is supposed to be. In "Thought and Talk," Davidson admits that

> At best what we have shown, or claimed, is that unless there is behavior that can be interpreted as speech, the evidence will not be adequate to justify the fine distinctions we are used to making in the attribution of thoughts. If we persist in attributing desires, beliefs, or other attitudes under these conditions, our attributions and consequent explanations of actions will be seriously underdetermined in that many alternative systems of attribution, many alternative explanations, will be equally justified by the available data (1975/1984, p. 164).

In his later article, however, Davidson seems to draw a much stronger conclusion from what is essentially the same argument:

> From what has been said about the dependence of beliefs on other beliefs, and of other propositional attitudes on beliefs, it is clear that a very complex pattern of behavior must be observed to justify the attribution of a single thought. Or, more accurately, there has to be good reason to believe there is such a complex pattern of behavior. *And unless there is actually such a complex pattern of behavior, there is no thought.* (1982b, p. 322, my italics).

The stronger conclusion put forward at the end of the second quote is clearly not warranted on the basis of the argument at hand. Davidson may in fact recognize this, since he is at pains to stress that the argument presented here is not his main argument. Perhaps he does not intend the italicized claim to be read as the conclusion of his first argument but rather as an

[5] Stich (1979) puts forward an argument along similar lines.

anticipation of the conclusion of his second argument. If so, the sentence is rather misleadingly placed.

Davidson is right to be restrained in his earlier appraisal of the argument. At best, what his argument shows is that we cannot be justified in attributing particular beliefs to animals, not that animals in fact lack beliefs entirely. Searle (1994) makes this point in his criticism of Davidson, and even Heil (1992), who wants to preserve as much of the Davidsonian picture as possible, feels compelled to admit this weakness. In addition to the obvious slip from "we cannot be justified in believing p" to "it is not the case that p," it is worth pointing out that it does not follow from the claim that we cannot ascribe any particular belief to an animal that we cannot justifiably claim of the animal that it has beliefs (though we know not which particular ones). To make this latter slip would be to act like the fellow who, when confronted with an ordinary gumball machine, reasons as follows: I can never be justified in thinking that a red gumball will come out of the machine (since only 25% of the gumballs are red), or in thinking that a green gumball will, or a blue one. Therefore, I can never be justified in thinking that the machine distributes gumballs at all. This fellow then walks away from the gumball machine, declaring it a waste of money. Davidson, if he means to draw the strong conclusion that animals do not have beliefs on the basis of the argument presented above, makes both the errors described.

However, even if Davidson were only right in his weaker claim that we could never be justified in attributing particular

beliefs to animals, that would be a major blow to those who wish to defend the idea of belief without language. Presumably, most of the defenders of this view would hold -- certainly I hold -- that we can in fact ascribe particular beliefs to creatures without language. Or, to put it more precisely, we can do so to some extent: humans and non-human animals are not qualitatively different in this regard. It is not the case that we can only be justified in attributing to animals only hopelessly rough, vague, and indeterminate beliefs -- beliefs without determinate aspectual shape -- while we can make human belief ascriptions with crystalline precision.[6]

Consider the following case. Mary, the owner of Ajax, is in the backyard with her dog and, like her dog, has observed the aforementioned cat. Imagine that we have learned from conversation with Mary that she is an avid hater of cats and is doing her best to encourage Ajax to chase them mightily so they will not plague her backyard. Now we have witnessed the cat running toward the oak, and we have witnessed its last-minute swerve up the maple. We see Ajax barking up the oak tree and clawing at its bark. We also see Mary peering up into the tree, pointing and saying, “Yes, Ajax! He went up that way! We’ll teach that trespassing pest never to enter our yard again, won’t we?” It seems quite natural to say that Mary, like Ajax, thinks the cat is in the tree.

[6] Dennett (1987), Routley (1981), and Smith (1982) each in different ways argue a similar point.

I wouldn't want to deny this. Notice, however, that the same kinds of questions may be raised here about Mary as were raised earlier about her pet. Is it better to characterize Mary's belief as a belief that the cat went up the tree or as a belief that an annoying pest went up the tree? Does Mary believe that the cat went up the only deciduous tree in her yard? Does she believe that it went up the only object on the block that was a sapling in 1908? Or that a creature who should not be in her backyard is probably higher up than it wants to be? Does she believe all these things, or just some of them, and which ones? And how can we tell? If we apply the same standards to Mary that Davidson wants us to apply to her dog, we may find ourselves committed to the position that *neither* of them has beliefs. In Mary's case, as in Ajax's, the evidence available to us is clearly not sufficient to warrant confidence about exactly what aspectual shapes her beliefs have regarding the events at hand. If Davidson requires that we withhold judgment about the content Ajax's beliefs on this basis, it seems we must also be forced to withhold judgment about the content Mary's beliefs.

It might be thought that there are crucial differences between Ajax and Mary that I have missed, which warrant us in ascribing particular beliefs in the one case but not in the other. One might argue, for instance, that Mary has the *concept* of a tree and Ajax does not, and that this difference is somehow key. I do not see this as a crucial difference for belief ascription, however, for two reasons. First, we often attribute

beliefs to people containing concepts they do not have, especially when those concepts are used to determine reference (as they are in the "cat" and "tree" cases here), or when the person has different concepts as a result of having a language that divides up the world in a different way. For instance, I might say of Paul that he thinks the man in the gabardine suit is a spy, even if I know that Paul has no idea what a gabardine suit is.[7] In a similar vein, then, why shouldn't I be able to say of Ajax that he thinks that the elegant Siamese we were just talking about is up in the tree, even if we grant that Ajax has no idea what a Siamese is or what it was we were talking about? In foreign language cases, also, we tend to find ourselves ascribing beliefs to people involving concepts they do not have. For example, I might attribute to an ancient Chinese philosopher the belief that a particular action is immoral, even though that philosopher might not have any concepts that match exactly with our concept of immorality -- the closest probably being *pu te* (not virtuous) or *pu yi* (not right).

Still, one might say, we wouldn't ascribe such a belief to a Chinese philosopher unless he had *some* concept approximately matching our concept of immorality. This brings me to my second point against the claim that the crucial difference between Mary and Ajax somehow turns on Mary's having the concept of a tree and Ajax's not having that concept. Even if we were to reject

[7] Such belief ascriptions are sometimes called *de re* belief ascriptions (e.g., by Quine 1966/1976). In *de re* belief ascriptions, there is a degree of semantic transparency. Roughly, a *de re* belief ascription may be cast in the form: *S* believes of *T* that *T* is (or has) *P*, where any means whatsoever can be used to pick out *T*, regardless of whether the person to whom the belief is ascribed considers *T* in those terms.

description of Ajax as believing that the cat is in the tree because he does not have the human concepts of cats and trees, that needn't mean that Ajax doesn't have concepts with similar extensions which function in a belief similar to the belief that the cat is in the tree, a belief approximately captured by that sentence. For example, Ajax might have a concept of a "tree*" as a tall thing with a shape something like this, leaves on top that sometimes come down, a smell something like this, and good for peeing on to mark territory. (Although again, such an English rendering can only be approximate: Ajax's concept of a leaf, and his concept of territory, are no doubt rather different from our own.) To insist without further argument that dogs cannot have beliefs of this sort begs the question against animal belief. To assert that a creature with a cluster of such beliefs still cannot have a *concept* of a "tree*" threatens to obfuscate the notion of 'concept' and render it useless to the debate. (If one attempted to define the word 'concept' in such a way that dogs could not have them, I would naturally question whether such things were really necessary for beliefs.) In any case, I don't see why having clusters of beliefs of this sort shouldn't be sufficient to satisfy Davidson's holism requirement mentioned above. A dog may know more about trees or snakes or bones (e.g. because he knows a lot about their smell and doggish uses, etc.) than many humans to whom we attribute beliefs about such things. Furthermore, given Davidson's holism about the content of beliefs -- his view that one's concept of a tree is the product of a wide

range of one's beliefs about trees -- *everyone* should have a slightly different concept of what a tree is. Perhaps I think a saguaro cactus is a tree and Mary doesn't. If this is true, then what I am doing when I say that Mary thinks the cat is in the tree is not different in kind from what I am doing when I ascribe Ajax the same belief: in both cases I am using an English sentence that only conveys *approximately* what I take to be going on in their heads. The difference is that in Mary's case, because our concepts and our worldviews are more alike, the approximation is a fair bit closer. (I will return to the issue of the approximate nature of belief ascription in chapters five through seven.)

A second difference between Mary and her dog is that we can question Mary about her beliefs. If we want to know whether Mary believed that the cat was in the only object on the block that was a sapling in 1908, we can ask her. It might be thought that this fact could serve as a starting point for an argument that we can ascribe particular beliefs to Mary but not to her dog. Imagine, however, the results of actually posing such a question. What kind of response are we likely to get? Clearly, if Mary doesn't know this fact about the tree she will deny having such a belief, but let's suppose she does recall -- now that we mention it -- that Aunt Janet planted the tree in 1906 in memory of her mother. In response to our query, then, perhaps we will get something like this: "No, I didn't believe that. Well, maybe I did. I don't know -- I wasn't really thinking about it that way

at the time. You philosophers ask such silly questions!" Even if Mary does come up with definite answers to our questions, we might wonder how much stock we ought to put in such answers. I am skeptical, then, about whether even in what might seem to be the most favorable cases, the cases in which we can ask a person directly about her beliefs, we can do what Davidson seems to want to require of us in the animal case: that is, nail down *specifically* what the content of Mary's beliefs is. For humans as well as for animals, our belief attributions will be seriously underdetermined by the available data.[8]

Perhaps we do know better what is going on in Mary's mind than in Ajax's (although I think this is an open question). If there is a difference here, however, it is only one of *degree*. We are not totally at a loss regarding how to describe Ajax's beliefs, nor are we capable of nailing down Mary's beliefs with spotless precision. Our efforts give us an understanding of dog and owner that lies somewhere between the two extremes. Some kinds of knowledge and ways of thinking about the world we know to be alien to Mary and her dog, some natural. We don't think Ajax considers the cat to be doing a dishonor to Grandma Szypanski's memory, nor do we think Mary likely to think of the cat in terms of its smell. We know something of the way Mary and Ajax approach the world and we can use our knowledge to provide us with a range of ways of approximating with language what we take to be going on in their heads. These epistemic facts

[8] Dennett (1987, p. 110-116) and Smith (1982) make a similar point. Note that although the point is an epistemic one, it seems to be employed by Davidson to make an

provide no basis for claiming an important ontological difference between the contents of Mary's mind and the contents of Ajax's. If Davidson continues to insist that there is an important ontological difference here, rooted in the greater "complexity" of language-users' behavior, he does so without a clear argument.

---

ontological point: There really is nothing specific to be nailed down.

2. Faults in Davidson's Second Argument Against Belief Without Language

It is clear in both his articles against the possibility of belief without language that Davidson attaches greater weight to a second argument than he does to the argument just presented. This second argument is quite simple and runs as follows (1982b, p. 324-327, 1975/1984, p. 169-170):

(P1.) In order to have beliefs, it is necessary to have the concept of belief.

(P2.) In order to have the concept of belief, one must have language.

(C.) Therefore, belief is not possible without language.

Granting that infants and animals are not capable of language, it follows immediately that they do not have beliefs. Unlike the first argument, this second argument is clearly valid. I will concentrate my attack on the first premise.

Both premises make reference to the "concept of belief." What does Davidson think this concept involves? In "Rational Animals" Davidson equates having the concept of belief with having a belief about a belief (1982b, p. 326). This may seem like too weak a requirement -- after all, one can have a belief about an ocelot without having the concept of an ocelot ("that cat looks so cute and tame"). However, Davidson glosses his claim in such a way as to make it clear that he means to be saying that the concept of belief requires the capacity to have beliefs about beliefs understood *as* beliefs. Although Davidson does not phrase his claim in this way, others have called the

capacity to which Davidson seems to be alluding "metarepresentation" (Heil 1992; Perner 1991b).

Davidson envisions at least two conditions that must be satisfied before he is willing to grant a creature the capacity in question:

(M1.) The creature must have the ability to recognize that a belief may be false.

(M2.) She must have an understanding of what Davidson calls the "objective-subjective contrast" -- i.e. the idea of "an objective reality independent of my belief" (1982b, p. 326, 1975/1984, p. 170).

It is interesting to note that the emergence of both of these capacities in children has been studied by developmental psychologists (e.g. Perner 1991b; Wimmer and Perner 1983; Gopnik and Astington 1988; Flavell, Green, and Flavell 1986), and they have been found to emerge at roughly the same time. If these psychologists are right, however, the abilities in question appear rather later than Davidson might hope: most children are *four years old* before they have these capacities. More on this shortly.

Assuming that the above is something like what Davidson has in mind when he mentions the "concept of belief" in (P1.) and (P2.), let's take a closer look at the plausibility of these premises. I intend to focus my argument on (P1.), but before doing so, I would like to look briefly at (P2.). Davidson claims that one cannot possibly have the concept of belief unless one

has language. In defense of this claim, Davidson confesses that he can offer only an analogy.

> If I were bolted to the earth I would have no way of determining the distance from me of many objects. I would know only they were on some line drawn from me toward them. I might interact successfully with objects, but I could have no way of giving content to the question where they were. Not being bolted down, I am free to triangulate. Our sense of objectivity is the consequence of another sort of triangulation, one that requires two creatures. Each interacts with an object, but what gives each the concept of the way things are objectively is the base line formed between the creatures by language. The fact that they share a concept of truth alone makes sense of the claim that they have beliefs, that they are able to assign objects a place in the public world (1982b, p. 327).

What Davidson says about physical triangulations is, I think, false: a person bolted to the earth could learn to mark distance by noting cases of occlusion and interaction and the relation of these to differences in the perceptual size of objects; furthermore, it is not clear that triangulation is the primary means people who are not bolted down use to judge distance. Of course, this doesn't prove false his remarks about "triangulation" by means of linguistic interaction between people. These rather cryptic remarks are the subject of substantial sympathetic decoding by John Heil (1992, p. 214-222). Heil suggests we understand the requirement of triangulation as a requirement that we be able to *compare* our view of the world with the view of another. Only if we are able to do this can we understand that our view of the world is just that -- a *view*. And this understanding is plausibly connected with requirements (M1.) and (M2.) above. But why is language necessary for all

this? At this point, Davidson would likely appeal to an idea he defends in "Belief and the Basis of Meaning" (1974/1984): Language is necessary for triangulation because we could not come to understand another's beliefs without simultaneously understanding her language.

I suspect Davidson could be fruitfully challenged regarding (P2.) and the triangulation metaphor. I have gone some way in the previous section, I hope, toward undermining his idea that we can't give content to the beliefs of a creature without language. Even Heil, though generally sympathetic to Davidson's project, has some qualms about (P2.). Heil describes various circumstances in which it might be possible for a creature without language to come to understand that her beliefs might be false, might not match up with the way the world actually is. Perhaps Heil is right about this. Nevertheless, I am willing to concede (P2.) for the sake of argument. I will argue below, in fact, that (M1.) and (M2.) emerge relatively late in the development of youngsters, well after the development of language, and I have never seen any convincing study suggesting that these capacities are present in non-human, non-language-speaking animals.[9] Maybe for some reason Heil missed language *is* necessary for the concept of belief. Davidson has not, I believe, presented a convincing argument in this direction; on the other hand, I have no argument against it.[10]

[9] Woodruff and Premack (1979) have a well-known argument for the existence of such capacities in chimpanzees, but there are substantial difficulties with this argument, difficulties admitted to by Premack himself (1988).

[10] Bishop (1980) also presents an interesting argument against (P2.).

Against (P1.), the claim that belief is impossible without the concept of belief, I am better prepared to argue. First, notice that Davidson's arguments in favor of (P1.) are rather limited. In "Thought and Talk" he says only this in defense of the premise:

> Can a creature have a belief if it does not have the concept of belief? It seems to me it cannot, and for this reason. Someone cannot have a belief unless he understands the possibility of being mistaken, and this requires grasping the contrast between truth and error -- true belief and false belief. But this contrast, I have argued, can emerge only in the context of interpretation, which alone forces us to the idea of an objective, public truth (1975/1984, p. 170).

The defense here amounts merely to a restatement of (P1.), not in terms of the concept of belief in general, but rather in terms of what Davidson regards as a requirement for having that concept -- the capacity to recognize that one's beliefs might be false (M1.). To this is added a restatement of (P2.). This defense, in other words, is no defense at all.

In "Rational Animals" Davidson does a little more in way of defending (P1.). His argument runs as follows (1982b, p. 326). I cannot have a belief unless I have the potential to be surprised. But surprise requires that I become aware of a contrast between what I did believe and what I came to believe. This requires a belief about a belief (understood *as* a belief): I came to believe that my original belief was false.

The argument, though perhaps initially attractive, does not stand up to scrutiny. It is not a necessary condition of surprise as we ordinarily understand it that one come to

recognize a past belief as false. I might be surprised to find that I have won the lottery, though I do not judge myself as having been earlier mistaken about my chances (or anything else). The argument thus turns upon a false premise, and second step in the argumentative chain from having beliefs, to having the capacity for surprise, to having the concept of belief, is cut. Davidson might wish to escape this objection by saying that he means something different by "surprise" than what we normally mean by it -- on Davidson's understanding of "surprise," perhaps, surprise entails the recognition of a past false belief. But then there would seem to be no reason to accept his claim that belief requires the capacity for surprise -- no reason, that is, unless we already accept (P1.). But (P1.) is supposed to be the *conclusion* of the argument, not a premise. Davidson's argument from surprise, then, is either question-begging or it rests upon a false premise. Either way, it provides no support for (P1.).

The simplest reason to reject Davidson's second argument, then, is this: it has a dubious first premise which Davidson gives us no good reason to accept. Why should having a belief require the concept of belief any more than having a pain or a bad temper requires the concept of pain or bad temper?

John Heil devotes considerable effort in his discussion of Davidson to making (P1.) seem plausible (1992, p. 198-205). Heil's argument is this. In some sense of "representation," many things may be thought to have representational properties. For instance, the bimetallic strip in a thermostat is a device

designed to represent temperature by curling to a greater or lesser extent depending on the temperature, closing the connection to the furnace when the air is too cold. In the natural world, honeybee dances may be thought to represent the location of honeybee food. But, Heil thinks, such representations to do not by themselves have determinate aspectual shape, as beliefs do; descriptions of the representations do not exhibit referential opacity.[11] There is no fact of the matter, Heil thinks, whether thermostats measure air temperature as opposed to mean kinetic energy of nearby molecules (or any like quantity) -- thermostats represent all such related quantities just the same. The case is similar for honeybee dances: can we really insist that the honeybee dance represents the location of food as opposed to the location of (say) a chemical substance of type F associated with the presence of food? With greater knowledge of honeybees, we may be able to rule out certain candidates in this department, but there will always be, Heil thinks, some important range of options, with no clear basis for our preferring to describe the honeybees as representing things one way rather than another.

Heil goes on to argue that it is only in a system with the capacity for *metarepresentation* that representations acquire definite aspectual shape.[12] (A "metarepresentation" is a

[11] Heil actually uses the term "semantic opacity" to talk about both the referential opacity of sentences and the fact that beliefs have aspectual shape. I think the application of such linguistic terminology to beliefs is apt to be misleading, so I will not follow him in this. I do not think my reinterpretation of Heil's terminology makes a difference to the argument at hand, however.

[12] It actually may be the case that Heil only wishes to argue that metarepresentation *suffices* for the possession of cognitive states with aspectual shape, rather than being *necessary* for it. I shall interpret him as making the necessity claim, since without it

representation of a representation understood *as* a representation. For the sake of argument, we can grant that a creature has this capacity just in case it satisfies (M1.) and (M2.) above.) Why is this? It is because metarepresentations, taking other representations as their content, are capable of exploiting differences in the aspectual shape of a representation in a way no other system in a creature can. Only if honeybees had the capacity for metarepresentation could a representation that there is a chemical F in a certain location generate different behavior from the representation that there is food in that location. And unless a creature can entertain representations with aspectual shape -- where representations with different aspectual shape have different impacts on behavior -- that creature has no beliefs.

Heil's argument is a difficult one, and I hope my presentation of it has been fair. I must admit I have trouble seeing the pull of it. First, I would like to reject the premise that only if a system is capable of exploiting aspectual shape behaviorally can it be said to have representations with aspectual shape. Heil (p. 198) cites Fred Dretske (1988) on representation as though he wishes to begin a Dretske-friendly discussion of representation -- and to a point what he says about representation *is* a lot like what Dretske has to say. But on Dretske's account of representation, an object represents what it has the function of indicating, and we can build a bimetallic

---

his argument cannot succeed as a defense of (P1.): unless metarepresentation is necessary for aspectual shape, the possession of beliefs will not imply the capacity for

strip with the function of indicating temperature *specifically* (as opposed to mean kinetic energy). Or -- to use an example less likely to run us into definitional and scientific problems -- we can (and generally do) build fuel gauges with the function of indicating the amount of fuel left in the gas tank as opposed to the amount of downward force exerted by the fuel tank on the bolts holding it to the car frame, despite the fact that the fuel gauge generally indicates both quantities (1988, p. 59-60). On Dretske's account, then, the representations my fuel gauge provides me with do have aspectual shape -- and claims about what my gauge is representing are referentially opaque -- despite the incapacity of the device to exploit this aspectuality in its behavior. If we try to make the case more analogous to the thermostat case by taking gauge-reading humans out of the picture -- perhaps by imagining the fuel gauge to have some control function in an automatized car -- the situation does not change. The gauge still has the function of indicating the amount of gas left. It does not malfunction if, for example, the vehicle is transported between the earth and moon so the gauge no longer reliably indicates the downward force exerted on the bolts.[13] Similarly, depending on one's account of natural functions, one might think there is good reason to say that the honeybee's dance represents the direction of *food* specifically, as opposed to the presence of chemical type F (or vice versa), despite the fact

---

metarepresentation.

[13] This, of course, must be done by a human representer; so only might argue that in a rather roundabout way metarepresentational capacity is presupposed even in this case of referential opacity.

that these two factors are generally correlated. (I will have more to say about representation in chapter three.)

A second crucial assumption Heil makes in his argument is that only if a creature has metarepresentational capacity can that creature exploit the aspectual shape of its representations. I am not sure exactly what work "exploit" is supposed to do here, but I suppose Heil's claim must amount to something like this: only if a creature has metarepresentational capacity can it make functional use of the fact that its representations have aspectual shape. It is a bit difficult to imagine what sort of functional use we make of the fact that our representations have aspectual shape. Examples meant to show that our representations have aspectual shape typically involve cases of ignorance or misrepresentation for which it is doubtful there is a specific function. I believe that Carl just came home, but I don't believe that the president of the bank just came home, despite the fact that Carl is president of the bank. How, exactly, am I supposed to "exploit" the aspectuality of this belief?

One case that does come to mind in which we might be said to exploit the aspectuality of our beliefs is in being prepared for counterfactual situations: I believe Carl came home and I know Carl is president of the bank, so I believe the president of the bank came home, but because these two beliefs are different beliefs with different aspectual shape (Heil says they are "fine-grained"), I could just as easily -- in a different possible world -- have believed one without believing the other.

But this means of exploiting the aspectual shape of representations is not confined to metarepresenters. Consider again our automatized car. Suppose this car has a fuel gauge whose function it is to indicate when the fuel falls below a certain level, so that the car can "report in" for refueling. Suppose also that it has another gauge whose function it is to indicate when the weight of the liquid in the gas tank falls below a certain level so that the car may take advantage of its lighter weight in maneuvering. Now, in fact, both these devices always go off at the same time. (The engineer who designed the gauge setup of the vehicle was fired for this blatant inefficiency.) But the car would be capable -- if the world were a different place -- of registering these two facts separately.

Perhaps I am missing something obvious in Heil's argument, but without a better sense of exactly what it means to be able to exploit the aspectual shape or "fine-grainedness" of representations, it is difficult to judge whether a creature or machine without metarepresentational capacity could do so. Even if Heil were right about this point, however, his argument could still be challenged on the grounds that it is not obvious, for reasons discussed above, that a creature without the capacity to exploit the aspectual shape of its representations would necessarily thereby *not* have representations with aspectual shape.

Do we have any reason, then, for accepting (P1.)? I think not. Neither Davidson's nor Heil's defense of this premise gets

off the ground. And on the face of it, (P1.) is not particularly appealing. It may be the case that in order to have a belief a creature must be able to distinguish, at least rougly, states of affairs that would count as satisfactions of that belief from states of affairs that would not -- perhaps we shouldn't be willing to say that Ajax can believe that Mary is home unless he can in some general way distinguish states of affairs in which Mary is home from states of affairs in which she is not -- but this is a far cry from having the metalinguistic notions of truth and falsity and the capacity to think of one's beliefs as possibly true or false (Searle 1994). Why anyone should think (M1.) and (M2.) necessary for belief is, I have to admit, something of a mystery to me.

There is a simple but important rebuttal to Davidson's argument, then. It is merely this: the argument depends on a counterintuitive premise for which neither Davidson nor his supporters are able to provide convincing support. There is simply no reason to accept (P1.). In the remainder of the section I shall focus on a second argument against Davidson which is quite a bit more complicated. But before heading into that argument, I wanted to pause for a moment to consider the weight of this simpler, and in some ways more appealing, first argument, which I dub the "huh?" argument, as in, "(P1.)? Huh?"

* * *

My second argument is also an attack on (P1.), but one with perhaps more force than merely showing that Davidson presents no good reason to accept (P1.). I argue that (P1.), given a few sensible auxiliaries, commits one to a position about the timing and development of linguistic and metarepresentational abilities -- a position that has been shown empirically to be false.

I have already mentioned the empirical finding I think causes trouble for Davidsonians: Children generally do not develop the concepts of objectivity and false belief until their are four years old, or so a number of developmental psychologists say (e.g., Perner 1991b; Flavell, Green, and Flavell 1986; Gopnik and Astington 1988). Yet most children are actively using language by the time they are two.

These findings should be troublesome for Davidson because he is committed to the position that language and the understanding of false belief and objectivity must emerge *simultaneously*. Obviously he accepts the claim that one cannot understand objectivity and false belief until one has language -- that is just (P2.). But he also thinks the conditional runs in the other direction. At the beginning of “Thought and Talk” Davidson says that “the dependence of speaking on thinking is evident, for to speak is to express thoughts” (1975/1984, p. 155). Indeed a project like radical interpretation (1973/1984) would make little sense if attempted on a creature without beliefs. But if speech requires belief and belief requires (M1.) and (M2.), then clearly speech must require (M1.) and (M2.). So the conditional runs

both ways for Davidson. Not only does an understanding of objectivity and false belief require speech, but the possession of speech requires an understanding of objectivity and false belief.

Therefore, unless Davidson wishes to claim that children are exempt from natural law and philosophical theorizing (a claim to which I admit I have sometimes been tempted), he must be committed to the position that the two capacities develop simultaneously. Otherwise, every child would, at some point, be a counterexample. But, in fact, language does *not* emerge at the same time in children as (M1.) and (M2.) do. It emerges much earlier. Davidson's position therefore must be false.

There is a limited range of alternative responses a Davidsonian could make to the charges I have just leveled. She could: (1.) challenge the merits of the empirical research in question, (2.) deny that Davidson's claims are empirical (and so are not empirically falsifiable), (3.) deny that children really have "language" until they are four or so, (4.) accept less stringent criteria for possession of the "concept of belief," or (5.) try to make a gradualist case, arguing that children have the beginnings of the concept of belief and the beginnings of language at two and develop the two in tandem until they are four years old. In the remainder of this section I will examine each of these potential responses in turn.

So how good is the empirical research I cite? It is fairly widely accepted in the developmental literature, and to the

extent there is disagreement, there are few who would locate the development of an understanding of objectivity and false belief as early as the second year, when language emerges.[14] The debate has primarily been between those who hold that such understanding doesn't emerge until around the fourth birthday and those who think it emerges sometime around the third birthday (e.g., Wellman 1990; Sullivan and Winner 1993). Of course, arguments from authority don't hold any water in philosophy in the twentieth century -- I certainly wouldn't accept such an argument -- so I will try to explain what the research has been and why I find it convincing. This will take a few pages.

Let's take (M1.) first, the ability to recognize that a belief may be false. A seminal study on the developmental emergence of this ability was conducted by Heinz Wimmer and Josef Perner (1983). In this study, Wimmer and Perner told children some simple, concrete stories which adults would judge to involve false beliefs, and then asked the children questions intended to reveal whether they, like adults, would judge the characters in the stories to have false beliefs. One such story ran as follows (experiment 2, abbreviated rendition taken from Perner 1991b):

> "Maxi and the Chocolate"
> Maxi is helping his mother to unpack the shopping bag. He puts the chocolate into the GREEN cupboard. Maxi remembers exactly where he put the chocolate so that he can come back later and get some. Then he leaves for the playground. In his absence his mother needs some chocolate. She takes the chocolate out of the GREEN cupboard and uses some of it for her cake. Then she puts it back not into the GREEN but into the BLUE cupboard. She leaves to get some eggs and Maxi returns from the playground, hungry.

[14] Alan Leslie (1988) is a possible exception.

Test Question: "Where will Maxi look for the chocolate?"[15]

This story was told not just verbally, but with the use of puppets and miniature cupboards, so the children could better focus on what was going on. It was hoped that children who understood the possibility of false belief and the conditions under which false beliefs were acquired would guess that Maxi would look in the green cupboard, and that children who did not recognize the possibility of false belief or who were confused about how false beliefs were acquired would guess that Maxi would look in the blue cupboard.

Young children performed quite poorly on this test, almost never guessing that Maxi would look in the green cupboard. Four and five year olds answered correctly about 50% of the time, with five year olds -- but not four year olds -- performing at ceiling if told that the question was tricky, and that they should "stop and think." Four year olds were helped substantially if the story was changed so that all the chocolate was used up in the cake, in which case the actual presence of the chocolate in the blue cupboard would not be a distraction to the recognition of the fact that Maxi would look in the green cupboard. Even in this last condition, however, the three year olds failed 85% of the time to guess correctly.

[15] Since this experiment was conducted in Salzburg, I presume that it was conducted in German and this is a translation. I suppose it is something of a question whether the capacities of German-speaking and English-speaking children might differ on such tasks. I have not seen any results which suggest that they do, and at least one study that suggests they do not (Perner, Leekam, and Wimmer 1987). Penny Vinden (1996) has found differences in the developmental timing of this capacity between children in our culture and those in certain pre-literate cultures, however.

The fact that children under four consistently failed these tests could not be explained by the failure of the children to understand words like ‘know,’ ‘believe,’ etc. because such words were not used in the experiment. Many three year olds did forget where Maxi originally put the chocolate, but the four year olds did not forget and still performed poorly; furthermore, in a similar experiment conducted later (Perner, Leekam, and Wimmer 1987), the great majority of three year olds did remember the relevant facts -- including an additional fact which was emphasized, that Maxi did not see his mother move the chocolate -- and their performance was still below 50%. (Young three year olds answered correctly 21% of the time, older three year olds 60% of the time.)

What might explain these results? One hypothesis that has been proposed is that the problem is not with recognizing the possibility of false belief, but rather with understanding the conditions under which false beliefs are formed (Wimmer, Hogrefe, and Sodian 1988; Leslie 1988). Another possibility is that children recognize that the characters in the stories have false beliefs, but don’t understand the connection between belief and action well enough to guess that the false beliefs will lead to unsuccessful actions. A third possibility is that there is some sort of linguistic failure: The children don’t understand the question, interpreting it, e.g., as a question about where the chocolate really is.

A variation by Gopnik and Astington (1988) of an experiment originally designed by Hogrefe, Wimmer, and Perner (1986)

suggests against the first two of these interpretations. In this experiment, children are presented with a typical container, for example a “Smarties” box (Smarties are a candy well-known to British and Canadian children), and are asked what is inside. Naturally, they answer, “Smarties.” The box is then opened and the children are shown that it really contains a pencil. In the original experiment, the container was reclosed and the children were asked to guess what their friend, waiting in a separate room, would think was in the Smarties box if it was shown to him all closed up. As suspected, the children tended not to predict a false belief -- they said their friend would think a pencil was in the box. In the Gopnik and Astington variation on the experiment, the children were inquired instead about their own previous belief: did they think, when they first saw the closed box, that there were Smarties in it, or did they think it contained a pencil? Amazingly enough, a majority of three year olds reported that they had thought the box contained a pencil. This result cannot be attributed to the children’s generally poor memory; they remember quite well when their past belief is a true one, when the Smarties are visibly replaced with a pencil. The result also cannot be explained by the children’s reluctance to admit their own past error; they do just as poorly when asked to report another child’s mistake (Wimmer and Hartl 1991).[16] In fact, Wimmer, faced with his own experimental evidence, was forced to recant his earlier position, cited above, that the best explanation of his and Perner’s 1983 experiments was not that the

[16] This experiment was conducted in German.

children misunderstood false belief but rather that they did not understand the conditions under which false beliefs were formed. In the task at hand, knowledge of how beliefs are formed is not necessary and cannot explain the children's failure.

The possibility that the children's poor performance may be due to linguistic failure is contravened by the the fact that experiments conducted using a wide variety of tasks and question-types have generally produced the same results. Some have not used *questions* at all, but simply motivated the children to deceive another person, though the results on these experiments have been more mixed (see Sullivan and Winner 1991, 1993; Sodian 1991; Sodian et al. 1991; Peskin 1989 reported in Perner 1991; Hala, Chandler, and Fritz 1991). Furthermore, even if there were systematic linguistic misunderstanding throughout this wide variety of tasks wouldn't the most natural explanation of the consistency of such misunderstandings be the children's failure to grasp the concepts being tested for?

These experiment, in conjunction with Wimmer and Perner's 1983 experiments, strongly suggest that children have difficulty understanding the concept of false belief before they are four years old, even to the point of misremembering recent events involving false beliefs. Gopnik (1990) compares this active misremembering with that of a person committed to a theory who misremembers an anomalous event in such a way that it conforms with her theory. (I will discuss children and theories in substantially more detail in my next chapter.) Viable

alternative explanations of these experiments and others like them have not been forthcoming.[17]

A second ability Davidson requires before he is willing to grant that a creature has the concept of belief is an understanding of the "objective-subjective contrast" (M2.). Davidson does not explain exactly what he thinks understanding this contrast involves, but I think it is fair to assume that it involves understanding at least

(M2*) Things can sometimes appear to be one way when really they are quite another.

A creature who did not understand (M2*), who did not understand the difference between appearance and reality, would necessarily not satisfy (M2.).[18]

The development of the understanding of (M2*) in young children has been studied extensively by John Flavell and his colleagues (for example in Flavell, Flavell, and Green 1983, 1989; Flavell, Green, and Flavell 1986; Flavell, Green, Wahl, and Flavell 1987). In one experiment (Flavell, Flavell, and Green 1983), Flavell showed three and four year old children a sponge that looked like a piece of granite. When they first saw it, nearly all the children said it was a rock. Then the

[17] Sullivan and Winner (1993) and Wellman (1990) have managed to elicit, under very particular conditions, correct responding to similar experiments in children in their early threes, but it is doubtful that such responses are indicative of a general understanding of false belief. And even if we were to take such experiments as revealing a real understanding of false belief, that still would not save Davidson's thesis, since the onset of language is much earlier, usually before the child's second birthday. Jerry Fodor (1992) is one who interprets Wellman's results as suggestive of real understanding, but even he, despite his nativist promptings, is not brave enough to attempt defense the view that the understanding of false belief emerges as early as the second year.

[18] Those interested in exploring the variety of meanings the term "objectivity" has taken in recent philosophy are directed to Elisabeth Lloyd (1995).

experimenter squeezed it and allowed the child to do so. The child was then asked two questions:

(A.) When you look at this with your eyes right now, does it *look like* a rock or does it *look like* a piece of sponge?

(B.) What is it *really, really?* Is it *really, really* a rock or is it *really, really* a piece of sponge?

The younger children did not perform very well on this kind of test, tending either to give "phenomenalist" answers to both questions (it looks like a rock and really is a rock), or "realist" answers to both questions (it looks like a sponge and really is a sponge). Similar results were found with stone eggs, red tiles moved behind sheets of plastic to look black, and many other objects (with different proportions of realist versus phenomenalist answers for different objects). In the vast majority of Flavell's experiments, three year olds tended to resist saying that things could *look* one way and *really be* another, suggesting a lack of understanding of (M2*) (and therefore (M2.)).[19] This resistance persisted despite efforts on Flavell's part to make the tasks and language as simple as possible, and even in the face of attempts to train the children in proper use of the distinction (Flavell, Green, and Flavell 1986; Flavell, Green, Wahl, and Flavell 1987). Interestingly, Gopnik and Astington (1988) found age-independent correlations

[19] One might object that perhaps in the child's worldview a sponge rock *is* really a rock, just an unusual kind of rock, and so in the example cited, it would be perfectly acceptable for the child to say both that it looks like a rock and really is one. This objection may be plausible for individual cases, but does not address the fact that across

between performance on these tasks and performance on the false-belief tasks described above.

There are a few difficulties, I think, with Flavell's experiments. For example, there may be linguistic difficulties for the children, interfering with their performance on the tasks. (Flavell tries to control for this in Flavell, Green, Wahl, and Flavell 1987, but I do not think he succeeds.[20]) Also, there are a few tasks on which the children *did* seem generally to be able to give the right answers, although these were only a small percentage of the total tasks Flavell reports and not unlike other tasks on which he reports failure (the most notable examples are in Flavell, Flavell, and Green 1983, exp. 1). Still, the overall weight and diversity of Flavell's tasks is impressive, as is the children's remarkable resistance to training.

Although Flavell's studies discussed here, and Wimmer's and Gopnik's discussed above, may not be completely impervious to the challenges of skeptics -- what study is? -- they are at least highly suggestive, and on the occasions they have been adapted in attempt to address the challenges of critics (for example, by changing the language or details of the tasks), they have continued to generate results similar to those cited here. For these reasons, I think Davidsonians have a tough road ahead of

---

a *wide range* of cases it is difficult to get children to distinguish between appearance and reality.

[20] If you read the experiment, compare the children's performance on the "semantically transparent" A-R task with their much better performance on the "Pieces 1" task, supposedly a control task. Why shouldn't the latter task be considered a better test of their ability than the more linguistically laden former task? In fact, the Pieces 1 task better matches Flavell's own description on p. 128-129 of how an appearance-reality test might be performed with minimal linguistic demands.

them if they wish to stake their philosophical position on the gamble that such studies are wildly mistaken -- so far off as to locate the development of capacities at four years which actually emerge during the second year.[21]

After this long diversion into empirical psychology, the reader may need reminding of where we stand. I have argued that Davidson's position that language and the concept of belief are mutually dependent commits him to a strong developmental thesis: that language and the capacities described by (M1.) and (M2.) emerge simultaneously. The empirical work I have cited suggests that this developmental thesis is untenable. If so, Davidson's position must be mistaken.

Above I outlined four responses, other than challenging the merits of the psychological work in question, that Davidson might muster against the charge that his view has been shown empirically to be false. I shall now briefly discuss each of these remaining four responses (numbered (2.)-(5.) above).

It is hard to see how the second response -- that Davidson's work is not empirical and so is immune to empirical refutation -- could possibly do. Although Davidson sometimes claims that his views are not empirical (e.g., in 1982b, p. 317), it is plainly the case that if Davidson holds language to be impossible without belief and thus without the concept of belief, then he must hold that there are no creatures who have language but do not have the concept of belief. This is a claim subject to empirical

[21] For an interesting, philosophically informed discussion of recent work in this area, the reader is directed to Perner (1991b).

examination. If it is found to be false, then *modus tollens* something in Davidson's original position must be mistaken. It is a matter of simple logic.

Even if we were to grant that Davidson's argument was wholly *a priori* (which it is not), we could still subject it to empirical examination. You can check a complicated addition problem, for example, by counting beads. If you get the wrong number of beads, you should re-examine your addition. If you know that you counted the beads right, then you know that your addition must have been wrong. For simple arithmetic problems, like two plus two, such empirical checking is pointless, but for complicated addition problems, it can be helpful (especially with an abacus or a calculator). Given that Davidson's argument, to the extent it is like an addition problem at all, is more like a complicated addition problem than a simple one, it is worth checking. It it fails empirically, it is flawed. Davidson cannot dispel an empirical objection, then, by saying that his argument is not an empirical one.

The third possible response, that children do not really have language until they are four years old, seems wild on the face of it. By the beginning of their second year, most children are already using their first words. By around eighteen months, they are speaking in two-word sentences, and not long after twenty-four months, they are using grammar productively -- using plurals and present progressives appropriately, and so forth, and speaking in full sentences. Three year olds are capable of

sustained and complicated conversations involving a wide variety of speech acts. Their grammar is not perfect, but I know no one who would want to equate poor grammar with complete lack of linguistic ability -- especially, I imagine, not the Davidson who wrote “A Nice Derangement of Epitaphs” (1986), an apology for the malapropism, defending the position that real linguistic communication can take place even when one party is hopelessly bad at lexical choice.

How could one possibly deny that three year olds have language? I do see one route by means of which Davidson could do this. In “Communication and Convention” (1985a) and “The Social Aspect of Language” (1991), Davidson endorses something like a Gricean (or Lewisian) position regarding the structure of intentions behind language (not that he agrees with Grice or Lewis in other respects):

> If communication succeeds, speaker and hearer must assign the same meaning to the speaker’s words. Further, as we have seen, the speaker must intend the hearer to interpret his words in the way the speaker intends, and he must have adequate reason to believe that the hearer will succeed in interpreting him as he intends. Both hearer and speaker must believe the speaker speaks with this intention, and so forth... (1985a, p. 22).

One might legitimately wonder whether a three year old could engage in so sophisticated a thought-process. Although Davidson is willing to allow that such intentions as are necessary for communication may not be (and normally are not) “consciously rehearsed” or “deliberately reasoned” (1991, p. 7), it may well be that three year olds are not even capable of *implicitly*

forming such complicated intentions (whatever that involves). (See Gomez 1994, however, for a defense of the view that, in some sense, they do form such intentions.) If complicated Gricean intentions are necessary for language use and if they are unavailable to three year olds, then plainly three year olds are not capable of language.

This would be a desperate route of escape for Davidson, I think. It seems much more sensible to deny the antecedent of the last conditional than to accept the consequent. Even if one did wish to lift Davidson out of the difficulty I have posed for him by claiming that three year olds are not capable of language, doing so would place Davidson in a new difficulty: he would have to say, of course, that they had no *beliefs* either. (That's the whole point!) This seems even a funnier thing to say than that they have no language. Alison Gopnik has remarked that it is difficult to tell from casual conversation with a four year old whether she will be able to pass the false belief and appearance-reality tasks. Are we to believe, then, that half of these children, superficially indistinguishable from each other, have beliefs and the other half don't? (Or, for that matter, that we are engaged in a linguistic exchange with half of them but not with the other half?)[22]

The fourth possibility I suggested as a response a Davidsonian might make to the empirical difficulty in question involves a revision of Davidson's criteria for the "concept of

[22] Alison Gopnik made this remark in response to a talk defending Davidson given by John Heil at Berkeley in spring of 1994.

belief." Perhaps if these criteria were suitably relaxed, emergence of the concept of belief in children could be made synchronous with the emergence of language. One candidate for such a criterion that suggests itself, perhaps because it has been studied so widely, is the appreciation of object permanence, first studied in depth by Piaget (1954). The development of an understanding of object permanence -- that is, the understanding that objects continue to exist even when they are not immediately being perceived -- seems to be a development closely tied to an understanding of the existence of an objective world. It is also a development that reaches fruition about the same time language use is getting started in earnest, around the middle of the second year (at least according to Piaget; but see Baillargeon 1987; Spelke et al. 1992). It is at this time, according to Piaget, that infants generally come to understand that most hidden objects exist *somewhere* and that systematic searching will generally pay off. Also, like language, development of the concept of object permanence has roots extending back into the first year. It is generally during the latter part of the first year that infants learn to search in a rather limited way for objects that have been hidden from them.

Another capacity that emerges at about the same time as language is the capacity for imaginative pretend play, the ability to treat an object or situation as something other than what it is known really to be (Piaget 1951). Perhaps, then, Davidson could avoid the charges of asynchrony by modifying his

criteria for a creature's having the "concept of belief" to something like (M1') that the creature has the capacity to engage in pretend play, and (M2') that she be able to recognize the continued existence of objects independent of her own perception.

Although some connections could clearly be drawn between (M1') and (M1.) the ability to recognize that a belief may be false, as well as between (M2') and (M2.) an understanding of the "subjective-objective" contrast," there would be some weaknesses in such a move. First, it is not clear anymore that what is being revealed warrants the title "the concept of belief" and so there is the risk that Davidson will lose his purchase on whatever intuitive appeal there might have been in the claim that belief requires the concept of belief. Second, and probably more important, the adoption of (M1') and (M2') looks *ad hoc;* it is not clear what the connection is supposed to be between these capacities and the capacity for language. Evidence suggests, in fact, that development in object permanence is not better correlated with development in linguistic ability than are other, apparently unrelated cognitive developments (Gopnik and Meltzoff 1993). Piaget has argued for a connection between the capacity for pretend play and the development of language: both, he thinks, require the capacity to regard items in the world as "symbols" (1951), but such an argument seems remote from Davidson's concerns and would require a substantial retooling of his arguments and positions.

The fifth and final proposal that might be offered on behalf of Davidson as defense against my empirical objection is perhaps the most sensible; yet at the same time, it is vague and unsatisfying and, like the previous proposal, rather *ad hoc*. It is this: Language and the concept of belief *do* emerge simultaneously. They both emerge slowly, starting during the second year and culminating in the fourth. That is, until the fourth year the child doesn't really fully have the capacity to use language, just as the child does not fully understand false belief and the appearance-reality distinction. Likewise, during the second year the child does have the beginnings of an understanding of false belief and the appearance-reality distinction, just as the child has the beginnings of language.

If this proposal is to be more than just a ploy, it has to be fleshed out to some degree. Perhaps the most promising avenue in this regard would be to incorporate parts of what I have said in the previous two proposals: The seeds of the concept of belief lie in the capacity for pretense and and understanding of object permanence, and the failure of three year olds to be fully linguistic consists in their incapacity to entertain complex Gricean intentions. Of course, more would have to be said here, and it would have to be hoped that development of the capacity to entertain Gricean thoughts is synchronous with (M1.) and (M2.), but the position is not absurd.

Still, the position is a strained one. To anyone not viewing development through the lens of Davidsonian theory, it must

certainly seem that a toddler's capacity for language far outstrips any understanding that toddler might have of the nature of belief. At 36 months, we find preschoolers saying such complex things as "You need to get your own ball if you want to play 'hit the tree'" and "When I grow up and I'm a baseball player, I'll have my baseball hat, and I'll put it on, and I'll play baseball" (Shatz 1994); yet at the same time these very same preschoolers are making the grossest, most naive errors on such simple-seeming tasks as those studied by Wimmer, Gopnik, and Flavell.[23] It is a strech to say of a child at 30-36 months either that she has the beginnings of an understanding of false belief or that she is not fully linguistic (and thus doesn't really have full-fledged beliefs); Davidson, if he is to take this route, must say both.

In this section I have argued against Davidson's second, more serious argument against the possibility of belief without language. The argument was divided into two premises: (P1.) that belief requires the concept of belief and (P2.) that a creature without language could not have the concept of belief. I was willing to grant (P2.), though I thought doubts could be raised about it, and focused my attack on (P1.) It was shown that Davidson provides no real defense of (P1.), and Heil's attempt to defend the premise on Davidson's behalf was found to be weak.

[23] Actually, these sentences are examples of speech from a toddler who previously displayed at least one instance of what would seem to be a recognition of the capacity for false belief (Shatz 1994, p. 160). Still, the sentences do not seem to be different in kind from sentences uttered by other three year olds who consistently fail on the false-belief and appearance-reality tasks.

Since (P1.) does not look independently plausible, its lack of argumentative support is a serious shortcoming. Furthermore, I presented reasons to think that (P1.) commits Davidson to a position that flies in the face of substantial empirical evidence from developmental psychology. I imagined five rebuttals Davidson might make to this empirical objection and undermined each in turn. In the final section of this chapter I shall present a positive argument on behalf of the possiblity of infant and animal belief.

Let me conclude this section with a speculation. A reaction several readers of Davidson have had to these sections is that I have missed Davidson's *real* argument against animal belief. The argument goes something like this: We have excellent reason to think that believing goes hand-in-hand with the interpretation of other speakers' utterances (see, e.g., Davidson 1973/1984, 1974/1984). But, obviously, creatures without language cannot interpret the utterances of others. Therefore, they can have no beliefs. Indeed, it does seem right to say that the rejection of infant and animal belief is a natural outcome of Davidson's system as a whole and its particular reliance on the idea of "radical interpretation"; and I would speculate that it is this relation, more than the arguments described in this chapter, that drives Davidson to his position on infant and animal belief. Why, then, does Davidson not appeal to this reason explicitly in his defense of the view that belief requires language? One reason suggests itself: Showing that his views on radical

interpretation imply that belief requires language does not show that belief actually does require language; one philosopher's *modus ponens*, it is sometimes said, is another's *modus tollens*. The reader might walk away more convinced that Davidson's views on radical interpretation are mistaken than that belief requires language. Therefore, Davidson's position is best bolstered by *independent* reasons for accepting the view that belief requires language -- and it is only to those reasons that he explicitly appeals.

For this chapter really to be complete, then, perhaps I should include a section treating Davidson's views on radical interpretation in which I both assess their plausibility and show their connection with the view that belief requires language. The reader, however, will be spared from this potentially long and arduous exercise. If Davidson chooses not to include such reasons explicitly among his defenses of the view that belief requires language, then I do not see that a person who is not interested in Davidson interpretation for its own sake should feel compelled to address those reasons in critiquing Davidson's articles: He appearently meant the articles to be free-standing. Furthermore, I would add that the task of interpreting Davidson's work on radical interpretation is no mean feat and would lead us quite far from the topic at hand. If the reader finds Davidson's work on this topic so compelling as to force the rejection of anything that contradicts it, I doubt there is anything I could

do, short of devoting my entire dissertation to the topic, that would have any chance of reversing her position on the matter.

3. The Word ‘Belief’

I have attacked Davidson on enough fronts, I hope, to convince the reader that his arguments against infant and animal belief are not compelling. This does not by itself, of course, show Davidson’s conclusion to be false. There might be a powerful argument Davidson missed. The conclusion might even (though right-thinking philosophers quail at the suggestion) be true despite a lack of any good argument at all on its behalf. The point of this section is to convince the reader that this is not the case.

For reasons discussed in the introduction to this chapter, I take the central question here to be a question about the use of the word ‘belief.’ The question is whether certain borderline uses of the word, picking out mental states of infants and animals, ought to count as correct and literal usage. Although one might think to treat this as a question about ordinary language, I set such considerations aside in this case for two reasons: (1.) I don’t think ordinary language yields a decisive answer to the question of whether infants and animals have beliefs (although certainly the sentence S *thinks* that p can be used in ordinary parlance to talk about the mental states of infants and animals, I don’t think the same is obviously true for S *believes* that p -- see Nelson 1983); and (2.) I think our purposes as cognitive psychologists and philosophers of mind may be sufficiently at variance with the purposes of ordinary users of English that the most helpful understanding of the term

‘belief’ for us may not match exactly with that of ordinary usage.

There are two techniques that are often used to resolve disputes about whether to include a borderline or disputed usage of a term as a correct and literal use. The first technique, probably the more familiar to philosophers, it to attempt to *define* the term in question, or supply necessary and sufficient conditions for its application, in such a way that it becomes clear whether literal use of the term would cover the case in question. Although in chapter six I shall attempt something like this for the word ‘belief,’ that will not be my approach here. Here I will pursue the second strategy of looking at our *purposes* in the use of the term and determining whether those purposes are well or poorly served by extension of the term to cover the disputed case in question.

To get something of a handle on how this might work for a word like ‘belief,’ consider a more mundane term like ‘restaurant.’ Ernie’s Bar has a kitchen in back from which patrons can order overpriced pizza, nachos, buffalo wings, and the like. Is it a restaurant? According to municipal code it is. It is subject to the taxation and regulation appropriate to restaurants, which is stricter than that applied to supermarkets and convenience stores which also sometimes sell prepared food. On the other hand, if a few of your friends were hungry and interested in going to a restaurant and you suggested Ernie’s Bar, they might respond, “that’s not really a restaurant.” Or if

you were to give your friend Angela directions to your new house, saying, “take Baker street off the freeway and turn right on the first block with a restaurant on the corner,” expecting her to turn right when she saw Ernie’s Bar, you’d be likely to get the poor woman lost (even if she knows that Ernie’s serves buffalo wings). Now we might imagine two philosophers debating the question of whether Ernie’s bar was *really* a restaurant. How might they resolve the question?

The debate shares a number of features with the debate over whether infants and animals have beliefs. Like the latter debate, the restaurant debate has both a linguistic and an empirical component. It can be cut into the two questions: (1.) what are the conditions under which it is true to say of an establishment that it is a restaurant? and (2.) does Ernie’s Bar in fact satisfy these conditions? If the disputants thought the second question was the point of contention, they might want to go out and see whether Ernie’s bar has separate tables, a full-time cook, and so forth. Let’s suppose, however, that in this case, like the infant and animal belief case, the dispute is not primarily an empirical one.[24] The disputants are both intimately acquainted with Ernie’s Bar. It is a dispute of the former sort, about what should properly be counted as a restaurant.

One thing the disputants might do, then, is analyze the term ‘restaurant’ in accord with our ordinary-language, pre-

[24] Of course, this is not to deny that empirical research might bear on the question of whether various creatures deemed borderline can be said to have beliefs, or even that on some analyses it might be an open empirical question whether infants and dogs have the capacities judged necessary for belief. As a matter of fact, however, people have tended to stay away from the latter sort of position (possible exception: Chater and Heyes 1994).

theoretical ideas about what institutions are restaurants. This is how philosophical disputes have often gone. The term is analyzed either into old-fashioned sets of necessary and sufficient conditions or into clusters of features thought to be more or less central to the "restaurant" concept. If their interest is in ordinary language analysis, the debate might stay at this level. If the disputants are open to the possibility of deviations from ordinary use (as I hope philosophers will be in discussing 'belief') they might begin to ask a second type of question: What is the point of classing together all these things we call 'restaurants' under that single term? Will the purposes that motivate this classification be better served if we include Ernie's bar among "restaurants" or not? At this point, it will become clear that for different purposes different classifications might be appropriate. If we are interested in talking about the class of institutions to which one might go with friends in search of a meal that might be an adequate substitute for a meal prepared at home, Ernie's Bar will not count as a restaurant. On the other hand, if we are interested in talking about retail establishments with kitchens that should meet specific health standards, Ernie's Bar may well count. This may explain why your friends have different intuitions than city regulators about whether Ernie's Bar is a restaurant. Only after the purposes in using the term are made clear, will it seem sensible to propose an analysis of it. But by then the debate might be resolved and an analysis unnecessary.

I will take such a pragmatic tack in my discussion of the concept of belief. I will argue that for most of the purposes philosophers of mind and cognitive psychologists have in using the term, it makes sense to include mental states of infants and animals in the category we identify by means of the term 'belief.' It will not, then, be necessary to propose a specific analysis of the word 'belief' to resolve the debate: On *any* sensible analysis of this term that is sensitive to the general purposes of philosophers of mind and cognitive psychologists it should turn out that infants and animals have beliefs. If a philosopher wishes to use the term for some specific purpose that mandates the exclusion of infants and animals as potential believers, that purpose ought to be made clear beforehand, and it ought to made clear that the understanding of belief invoked is intended to be useful only within a specific restricted domain of inquiry and not across philosophy of mind and cognitive psychology generally.

The position, then, is a strong one. It is not to be confused with the much weaker claim that, whatever the reality behind the behavior we see is, it is convenient to treat infants and animals as though they had beliefs.[25] On my view, infants and animals *really do* have beliefs, supposing 'belief' in this sentence to be given the sense I endorse. And not only do I hold this, but I also think that on *any* general-purpose analysis of belief one wishes to propose for philosophers of mind and

[25] This position is often associated with Daniel Dennett (1987), although he may not be as anti-realist as he sometimes appears (see his 1991b for a discussion of this).

cognitive psychologists, one must be willing to grant that infants and animals have beliefs. In this sense, my position about animals and 'belief' is different from my position about Ernie's Bar and 'restaurant,' since in the latter case I did not see the preponderance of purpose weighing so heavily on one side of the question.

I am assuming for the argument that we are all philosophers of mind and cognitive psychologists here, interested in the word 'belief' because we think it plays a role in a helpful folk psychology and can be imported without serious damage into a sensible scientific psychology. As such, we feel free in a scientific or philosophical mode, if the evidence is right, to say of a creature that it has some belief or other. Abstracting away from (admittedly important) interpersonal, political, and other such situationally variable factors, I think our purposes in doing so are two:

(G1.) We want to predict and explain a creature's *behavior*.

(G2.) We want to predict and explain that creature's *phenomenology*.

On my view, the purposes described in (G1.) and (G2.) are happily met if we extend our belief ascription practices to cover infants, apes, and dogs. If so, then unless there is some other overriding purpose that gains our devotion, there will be no good reason not to count such an extension as a literal and correct use of the term 'belief.' We are, after all, making a practical decision about where to draw our lines.

Let’s look at our behaviorally based reasons for belief ascription (G1.) first. Think about Kim’s cat, Baby. Every evening, Baby hears the can-opener and food is placed in her dish. Today Baby has not eaten since morning. Now it is evening and Baby has a drive or desire -- or disposition, if you prefer -- to eat cat food. Suddenly, she hears the can-opener! Baby runs into the kitchen where her food dish is. A behaviorist might say that what we have here is a simple case of operant conditioning. Certainly there are examples of more complex cognitive processing in cats than this. Yet notice that it is perfectly natural to describe Baby’s behavior as caused, in part, by a mental state with many of the outward features of belief. As a result of an auditory perception of the operation of the can-opener, Baby’s brain shifted into a state which, because of the presence of a certain drive or desire, or at minimum a certain kind of disposition, resulted in behavior sensitive to the way things were in the world. This behavior will cause in turn the satisfaction of Baby’s drive or desire for food, or the instantiation and resultant slaking of her disposition to eat. Considering the plethora of similar examples in Baby’s life, we may with justice conclude that Baby has brain states that are belief-like in *at least* the following respects.

a. They may be caused by perceptual events.

b. They work in conjunction with desire-like states to produce behavior.

c. This behavior is sensitive to the state of the world, which is to say it would normally be different if the world were in a relevantly different state.

d. The states can "get it wrong" about the world (for instance, in cases of misperception) with the result that they generate inappropriate behavior. In this sense, we can say that these states have a "mind-to-world direction of fit" (Searle 1983), or that they are "representational" (Dretske 1988, 1993).

e. These belief-like states sometimes work productively together with other belief-like states to produce behavior that could not result from either belief-like state working alone. (Example: Baby sees Puddles, an enemy cat, lying in the path between her and her food dish, so she takes an alternate, roundabout route to the dish.)

f. These states have what I (following Searle 1992) have called "aspectual shape." I argued for this point in the first section of this chapter.

We have here a sizable array of behavior-related similarities between Baby's belief-like mental states and the beliefs of adult humans. If our interest is in behavior, on what basis might we be motivated to nonetheless deny that what Baby has are "really" beliefs after all?[26] There must be some crucial respect in which the relations between Baby's mental states and her behavior

differ from those of adult humans such that description of Baby's cognitive states as "beliefs" just isn't warranted.

One candidate that may suggest itself is expressibility. Baby cannot express her beliefs in language; adult human beings can. But what exactly are we to make of this? The condition that a creature cannot believe that p unless that creature can express its belief that p seems plainly too strong. I believe that my mother is Dutch, but that belief might cause in me so much distress that any time I try to express the belief, I faint halfway through. On a more mundane level, I might have a belief about exactly what shade of tangerine my new Volvo is without the verbal or artistic capacity to express this belief. Even the weaker claim that a creature cannot believe that p unless it can express *some belief or other* seems too strong. A car accident might cause my total paralysis, wiping out my capacity to express any of my beliefs, without thereby wiping out the beliefs themselves. Furthermore, it is just not clear why the capacity for expression in either the weaker or the stronger sense (or whatever other sense you wish to make of it) should be given decisive weight in the question of whether we should apply the word 'belief' to the mental states of a creature.[27]

I hope it is plain enough that if all we want is a model, not necessarily accepted with any strong accompaniment of realism, for the prediction and explanation of behavior, then a belief-desire model of mental content will serve us handily. As Dennett

[26] One might say that their mental states are "not propositional" -- but this is merely empty jargon unless it is cashed out in some way relevant to our purposes in belief

(1987) has suggested, if that is all we want, we can even often get away with ascribing beliefs to home computers. People who ascribe beliefs and desires to infants and animals do not thereby go vastly wrong in predicting their behavior. Hearne (1982) even suggests that people (mostly academics) who do *not* see animals as creatures with beliefs and desires tend to fail in training their pets and in predicting their behavior. Certainly, according to Hearne, most professional animal trainers work with models of animal cognition which closely parallel their models of human cognition.

However, even if we confine our purposes in belief ascription entirely to the explanation of behavior, we may want to occupy more of a realist position about belief than that described above. We might -- depending on our philosophy of science -- hold that a good explanation of behavior must appeal to mechanisms that not only generate the right predictions, but also are the mechanisms *really at work* in the mind. We want to tell the truth. Thus, we may want to extend our base of evidence beyond the merely behavioral to include the biological. (If there is any kind of evidence regarding the mental states of creatures beyond the behavioral and biological, it escapes me.) We may also want to include some discussion of phenomenology, grounded in behavioral and biological evidence. This latter subject I will pick up shortly.

---

ascription.

[27] McGinn (1982) makes a similar point.

Do we have, then, any biological reason to draw a fundamental line between explanations of adult human behavior and the behavior of the more interesting non-linguistic creatures? I think not. Perhaps someday we will have a biology capable of informing us about exactly what features of *homo sapiens* are responsible for their capacity to form beliefs. Such knowledge may -- or may not -- allow us confidently to distinguish the creatures capable of belief from those that are not and from those that are borderline in various respects. Our biology today tells us nothing so rich. As far as I can tell, our biological knowledge about belief is mainly this: Our brains are somehow centally involved in it. We can associate some of the larger regions of the brain with a few specific cognitive capacities, although this work has not come very far yet. We might even be willing to speculate that the parts of the brain that are evolutionarily the oldest, such as the brain stem, are not by themselves sufficient for the formation of anything we would want to call a belief. More than this we really cannot say. And of course babies, apes, and dogs have brains with much of the same gross structure as our own brains, and certainly much more to them than just a stem. For all we know biologically, then, the brain works the same way for them as it does for us: (in part) by harboring beliefs. Biology pulls more in favor of infant and animal belief than against it. One might even think that it creates a (defeasible) presumption in favor of animal belief.[28]

[28] Of course, one might say that the fact that we have language and these other creatures does not shows that there are some important biological differences among us --

To sum: One of our primary purposes in describing creatures as having “beliefs” is to predict and explain their behavior. I have argued that non-linguistic creatures can have mental states with a substantial array of belief-like features. If we treat these states as “beliefs,” incorporating them into a belief-desire psychology of the creatures in question, we do well in predicting and controlling the behavior of these creatures. Furthermore, we have no more biological basis to doubt that our predictive and explanatory success is the result of the creatures’ “really having” beliefs than we do in the human case. I conclude that if we wish to deny the practical virtue of having a notion of belief that covers infants and the higher mammals, it cannot be because our ordinary purposes in explaining behavior demand it.

What about the other purpose I described, the one with the phenomenological cast? Do animals and prelinguistic infants have mental states that play a belief-like role in their phenomenology? (By “phenomenology” here I mean something like subjective, first-person experience -- what things are like “from the inside” for the creature undergoing the experiences.) It might seem hard to know exactly what would count as conclusive evidence for or against this claim. We appear to be plunging into a domain from which a certain skeptical ghost has never quite been vanquished, the one that whispers in our ears that it is impossible to know of the existence or nature of “other

---

differences, perhaps, large enough to warrant belief ascription in one case but not in the other. The plausibility of this argument, however, seems to depend on the prior

minds." Even, perhaps, if we are willing to set aside such skeptical worries in the case of other adult human beings -- we think our neighbor Jocko Leibowitz must have subjective experiences that in important ways resemble our own -- we might think it rash to bring on board relatively more alien creatures like infants and dogs.

But why? It is plausible to think our phenomenal experiences are the product of our having brains of a certain type. Dogs and newborns also have brains -- brains, in fact, very much like our own -- so why not grant that they, too, may plausibly be thought to have phenomenology? Certainly there are differences between their brains and ours, but to hold that it is exactly those differences that are responsible for our having phenomenal, subjective experience, and that other creatures lacking these crucial brain features have no phenomenology at all, is a piece of speculative neurobiology that sounds suspiciously like an attempt to save a troubled theory.

It *looks* for all the world like infants and dogs have phenomenal experiences. They engage in behavior which, if analogs were found in any adult, would draw us unhesitatingly to the conclusion that there was phenomenology playing beneath. A dog sniffs up close to a raccoon and gets swiped across the nose. He yelps, leaps in the air, and runs away. He whines and attends to his nose. He is careful not to brush it against things for a while, and the next time he sees the raccoon he keeps his

---

acceptance of a tight connection between language and belief. It is no *independent* reason to think that animals without language cannot have beliefs.

distance. Who but a philosopher would deny that we was in pain? A baby who has not been fed since morning emits a certain high-pitched squeal that her mother has come to associate with the desire to be fed. The baby squeals continuously for a time with no obvious external cause, and upon seeing her mother increases the volume somewhat, for a duration. Upon being presented with her mother's breast, the baby relaxes and begins to feed voraciously. Who but a philosopher would say that this baby didn't feel hunger?

So I hope it will not be thought that I am assuming too much if I accept that infants and dogs have a phenomenology of sorts. At the very least, they can subjectively experience pain, hunger, warmth, loud noises, and so forth. Descartes was alleged to have kicked a cat while asserting that animals are really nothing but machines designed to squeak and make noise but soulless and so incapable of the subjective experience of pain (or anything else), but I do not think most skeptics about animal belief today would follow Descartes this far. Infants and animals may have phenomenology alright, but just not phenomenology of the right sort -- not the kind of phenomenology associated with genuine, honest-to-John *belief*. (Alternatively, the skeptic about animal belief might deny that the phenomenology is the important thing -- but then he'd have to rely on behavioral differences to do the work.)

It is worth pausing for a moment, then, to consider what kind of phenomenology *is* associated with belief. One piece of

phenomenology that might be thought to be rather central to believing is clearly not available to creatures without language: the experience of entertaining a verbal image in one's head and, in some sense, feeling assent toward it. An infant cannot say to itself, "I believe that Sesame Street will continue to attract a wide audience of young children" or even "Gee, that milk was nice and warm." An infant cannot express her beliefs in this explicit, verbal way. If one wishes to hold that this capacity is a *sine qua non* of belief, then it follows directly that infants and animals have no beliefs. No elaborate argumentation is necessary -- except, of course, to convince us to adopt the premise that belief requires the capacity to entertain verbal images.

On the face of it, it doesn't seem very plausible that belief requires that capacity. Consider, again, my brand new tangerine Volvo. What color, exactly, do I think it is? I do have a belief about its exact color. I would be surprised were I to go outside and find the car to be some *different* shade of tangerine. But no way can I express this belief verbally or entertain it as a verbal thought. And although somewhere deep down I understand that my mother is Dutch, I am completely incapable of entertaining a verbal representation of this fact -- it's just too traumatic for me. There are many instances of beliefs we cannot express with verbal images.

Although I wouldn't want to hang too much on it, an interesting case is described by André Lecours and Yves Joanette

(1980). These two psychologists studied an epileptic French monk ("Brother John") who, despite being on anti-seizure medication, was apt to have fits of "paroxysmal aphasia" which enormously impaired his capacity for the production and comprehension of language. Brother John reported proportional difficulty with inner speech. Although he claimed to be able to "think clearly," he was apparently unable to render those thoughts in words, even to himself. One interesting episode related by Lecours and Joanette is the following.

Brother John was travelling to Switzerland by train when he found himself at the height of an aphasic episode. He had never before been to the town that was his destination, but he had considered before the spell became severe that he was to disembark at the next stop of the train. When the train halted, he got off, recovered his luggage, and went in search of a hotel. Although presumably unable to read signs, he chose a building he judged likely to be a hotel and showed the person at the registration desk his medic-alert bracelet. When the person indicated by gesture that the hotel was full, Brother John sought and found another hotel and again showed his bracelet. He was able to provide the clerk the information necessary to complete a room reservation by showing her his passport, and was led to his room. Feeling depressed, he went downstairs in search of a snack at the hotel's restaurant, which he found by himself. Upon being given a menu, he pointed at what he hoped to be the desserts section, and was disappointed when the waiter brought him fish. After the meal he returned to his room and went to bed to sleep

off his fit. He awoke feeling embarrassed and felt the need to explain himself to the registration clerk, who apparently lent a sympathetic ear.

Of course, it is difficult to know how much credence to give to Brother John's self-reports about his incapacity with language during these aphasic episodes; and even if we do give Brother John full credence, his reported aphasia, though severe, was not complete -- we was *sometimes* able to match words to objects (but certainly not entire multi-word sentences). Nonetheless, it seems plain that during these aphasic bouts Brother John's capacity for intelligent action far outran his capacity with language. Furthermore, and of course more centrally for my purposes, it seems unnatural and unhelpful to deny him the capacity for belief during these episodes.

Another potentially interesting source of examples, which I would like someday to explore, would be studies of deaf people without sign language. I suspect their stories would not differ greatly from that of Brother John. I am not sure, however, to what extent such people could be granted a capacity for "language." My guess would be that these people would create stylized gestures by means of which they could communicate to a limited extent with those familiar to them. Whether such stylized communication, if it indeed occurs, should be termed "language" I am unprepared to say. If not, then we have an example of a whole range of adult human beings who are, unlike Brother John, *continuously* incapable of language. Even if we

want to grant that such people do have language, we may want to allow the possibility that certain deaf people, for some reason or other, never learn such a stylized repertoire of gestures. It would be empirically irresponsible, I think, (yet all too typical an example of philosophical hubris) to deny *a priori* that such people could be capable of a phenomenology which looks for all practical purposes like the phenomenology of belief -- except, of course, that it is accompanied by no verbal images.

We still haven't settled exactly what the phenomenology of belief is supposed to be. I have argued, or at least suggested, that it does not essentially involve the actual or potential presence of a verbal image, something uttered in an internal voice. Although it is not essential to my argument, let me go further and suggest what might seem to some a rather wild position: that belief, considered by itself, *has no phenomenology at all*. Certainly it is true that we have at all times a vast number of beliefs with no immediately present phenomenology. I say to myself now, "I believe Carter was President of the U.S. in 1978." I have had this belief since 1978, but it has not impinged constantly on my consciousness since then. For most of the time that has elapsed since 1978, this belief has occupied my head quietly, with no obvious phenomenal traces.

But, one might suggest, now that I am thinking of it, surely, my belief has a phenomenology! Well, what would this phenomenology be, exactly? I run a certain verbal image through my head -- I say to myself, "Carter was President in 1978" -- and

I feel some sort of assent or agreement with this claim. This verbal image and the feeling of assent accompanying it do indeed have a certain phenomenal character. But surely it is not *these* that constitute my belief. They don't have the right properties. The image and feeling come and go; the belief stays. The image has a particular English structure; the belief is independent of the exact form in which it is expressed (i.e. it is the same belief as that expressed by "In 1978, Carter was President"). Calling forth the image requires an act of will (albeit not a profound one); having the belief does not. Thus, the image and feeling, though they have a phenomenal character, are not the belief. But I can discover nothing else in the phenomenology of belief.

Having the belief no doubt *caused* me, in this circumstance, to entertain the verbal image and feel assent toward it (or perhaps the verbal image and assent are manifestations of a disposition which is the belief). It may also cause me, in other circumstances, to feel surprise (if, for instance, I were to find out that by some technicality of law Jimmy Carter's brother Billy was actually president in 1978). Beliefs, of course, play an important role in the generation of a wide variety of phenomenal experiences. I feel *anticipation* and *excitement* at the thought of that beer in the fridge I am about to drink, I *expect* it to taste a certain way, and I form an *image* of what it will taste like going down. I am *afraid* that it will explode when I open it up, since I just saw my roommate shaking it. I feel *disappointed*

and *angry* upon realizing that there is no way to drink the beer and keep my clothes clean at the same time. In my view, the role belief plays in phenomenology is its role in the production, behind the scenes as it were, of such images, feelings, and emotions.

We should ask, then, whether creatures without language can have such images, feelings, and emotions. The view that these creatures have beliefs commits one to the view that, by and large, these creatures do have this kind of phenomenology, and that their beliefs play a role in generating it.

I hope this will not seem implausible. If we are willing to grant, as I think we should, that infants and dogs have *some sort* of subjective, phenomenal experience, then I think we must grant that it goes beyond the pure sensations of hunger, pain, sound, and the like, but also includes feelings and emotions of various sorts. Obviously, some emotions are beyond the capacity of infants and animals -- I doubt an infant could feel wounded honor, for example -- but a basic emotional structure with various colors of positive and negative affect, at least, is surely present. And equally clearly, the emotions do not come and go at random but are affected by mental states with something of the look of beliefs. The baby becomes upset as a result of the mental state she enters upon hearing her mother leave the room. Ajax gets excited as a result of a mental state he enters seeing Mary reach for the leash. Brother John, if he counts during his aphasic episodes as a creature without language, is

disappointed when he sees that the waiter has brought him fish instead of a dessert.

Do these creatures also form nonverbal *images* something like those found in adult humans? The case is perhaps a bit harder to make here, but two considerations may come to the rescue. First, non-linguistic creatures plainly have the capacity to remember past events. If we grant that these memories have some phenomenal cast, it seems reasonable to conclude that they are imagistic. Second, there are scattered reports of "insightful" problem solving by primates that seem to require a capacity not only to entertain mental images, but also to manipulate them creatively. For example, a primate suddenly joins two short sticks together to make a longer stick that can be used to haul in a banana out of reach by means of either stick alone (Köhler 1926).

If our purpose, then, it ascribing beliefs to adults is to say something about how certain of their mental states relate to their phenomenology, that purpose may also be served if we choose to bring infants and dogs within the compass of the term. The latter, it would seem, also have mental states that play a belief-like role in the production of their phenomenology. Their phenomenology may be more limited in some ways, but so long as we are not tempted by application of the word "belief" to grant them a phenomenology beyond them (e.g. of honor or verbal images), then it seems that the extension of the term to these cases is perfectly natural, and a help.

Here, then, is a review of the argument so far. It was asserted, I hope plausibly, that the practical interests of philosophy of mind and cognitive psychology in belief ascription are primarily two. We are interested predicting and explaining behavior and phenomenology. Our purposes, therefore, in calling a state a 'belief' will be well-served if we call those things 'beliefs' that relate in the right way to these two aspects of a creature's life (or, at the very least, are members of a class of things most of which relate in the right way to the phenomenology and behavior of creatures). I argued, first, that some non-linguistic creatures have mental states with belief-like relations to behavior. These states have many of the formal and causal properties of genuine adult human belief, are grounded in a similar biology, and may be treated as beliefs for the effective prediction and manipulation of behavior. From a purely behavioral standpoint, it seemed that there was no good reason, then, to deny extension of the term "belief" to the mental states of non-linguistic creatures. Likewise, I argued that there is no good phenomenological basis to deny the extension of the term to non-linguistic creatures. Given that we grant (on biological and behavioral grounds) that such creatures do in fact have phenomenal experience, it is natural to suppose that this experience is not merely perceptual but also involves emotions and images. If it does, then it looks like the same states that play a belief-like role in behavior have a belief-like bearing on emotions and images. I argued that the ability to entertain verbal images is not necessary for belief. I saw no distinctive

phenomenology of belief apart from its role in producing images and feelings on the basis of which it would make sense to deny beliefs to non-linguistic creatures.

In the word ‘belief’ we have a useful tool for describing the mental states of creatures efficiently, with a broad range of behavioral and phenomenal implications packed in. If I tell you that Mary believes there is a cat up in that tree, you will be warranted in drawing a number of conclusions. You know that Mary hates cats, so you figure she will be upset. You figure she will probably go out in the backyard and take the opportunity to “teach the cat a lesson.” You figure that in her attempt to do so, she will approach the tree in question. And so forth. The word ‘belief,’ if used to describe the mental states of Ajax, supplies similar inferential power. If I tell you Ajax believes there is a cat up in that tree, you may then predict that he will be barking excitedly up into the tree and clawing at its trunk, he will be trying to detect any attempt on the part of the cat to escape, he is prepared to give further chase if necessary, he is probably all wound up and, given his rambunctious nature, it will probably require at least fifteen minutes for him to calm down. Our hammer seems to work as well on eight penny nails as it does on ten penny nails, so why should we use it only for the second job? This question gains special point when we don’t seem to have anything else in our toolbox that works nearly as well on eight penny nails as that hammer.

It seems to me the advantages weigh heavily in favor of giving the word “belief” a broad meaning, including infants and

animals in our belief talk. Perhaps the most decisive consideration in favor of this approach is just that we don't really have the tools to speak fluently about the mental lives of such intelligent but non-linguistic creatures *without* attributing them beliefs or the other intentional states normally attached thereto. Those who attack the idea of animal belief offer no helpful resources. Suppose we deny that Spot believes the neighbor's cat is in the tree as he stands, clawing at the trunk and barking into the branches above. Certainly Spot is in some mental state regarding that cat and the tree. What would that state be? How are we to describe it? Will we be forced back into behaviorist language and/or neuro-speak?

There are some alternatives. We might wish to retain *most* of the folk psychological apparatus, discarding only belief (and maybe one or two other terms considered inappropriate). Perhaps, though 'belief' is taboo, we can talk about what the infant or animal perceives and expects, what her concepts are, even, maybe, what she "knows" innately about the world. I'm not sure this kind of strategy makes a lot of sense. Can a creature expect or know something about the world, or have concepts, without having beliefs? Why single out belief for rejection? And if belief and desire are crucial elements of our folk psychological explanations, as they often are said to be, are we to abandon all such explanations -- or are "expectation" and desire explanations somehow better? If we are going to give folk psychology any

reign at all in our talk about infants and animals, it seems we have to let ourselves talk about beliefs.[29]

Another possibility, if we want to talk about the cognition of infants and animals without invoking the concept of belief, is to invoke computer analogies, quite popular these days. If we are serious and purist about our computer analogies, however, and see *adult* brains also as essentially big computers, and we think the same about animal and baby brains, why not grant that animals and babies have beliefs as adults do? If, on the other hand, we just want to use computer analogies as a way to get around talking about baby beliefs and we don't think adult human brains are really big computers, then we have committed ourselves to the unlikely position that babies, cognitively, are more like computers than like adult humans.

Other means of talking about infant and animal cognition without attributing beliefs to them include (1.) actually using the word 'belief' to describe what's going on in their heads but insisting continually that such use is metaphorical, or (2.) introducing a completely new set of terms, meant to apply specifically to the cognition of large-brained, intelligent creatures without language. I trust it is obvious why the second strategy has not been widely pursued. The first strategy, if taken seriously, collapses into an unclear version of the second: if the word 'belief' is to be consistently given two different readings, wouldn't it just make more sense to employ a different word and so avoid ambiguities? A third strategy would be to

[29] See also Routley (1981).

introduce a new taxonomy of mental states, either with or without roots in folk psychology, meant to apply both to humans and animals. I take it that this is what the "eliminativists" in philosophy of mind, for example, would like to do (Churchland 1981; Stich 1983). Although I am not opposed to such an ambitious project, we seem a long way off from being able to pull it off successfully.

If we take folk psychology seriously, as I have been doing, then we must grant that beliefs play a central, fundamental role in our cognition. To deny, then, that a creature has beliefs carries with it the suggestion that the creature's cognition, lacking this crucial element, has a radically different structure from our own. An immense gulf yawns open, dividing creatures capable of belief from those incapable of it, and we find ourselves standing alone on one side. Not only does this seem a mischaracterization of affairs, but it is one with potential moral consequences: If infants and animals are seen as so alien to us as not even to share the fundamental elements of our cognitive processing, might it not be tempting to accord their interests and welfare less weight than if we saw them as closer kin? Surely it does not follow as a matter of logic that those who wish to deny beliefs to these creatures hold them in less esteem -- I know at least one Davidsonian I am sure is an excellent parent! -- but it would not be surprising, I think, to find a correlation between the degree of regard in which a person holds such creatures and the degree of similarity that person finds between the creatures' cognition and her own.

Certainly there *is* a great divide between the cognition of creatures like ourselves and the "cognition," if we want to call it that, of creature with what might be more aptly called a cluster of nerve cells than a proper brain. To use a pre-Darwinian metaphor, we might think of such creatures as lying at the far end of a smooth and gentle spectrum proceding upward by imperceptible degrees toward humanity. At what point along this spectrum the capacity for phenomenal experience appears, and if it appears suddenly or fades in by degrees, I am not prepared to say. But it seems to me that the act of withholding the word "belief" from description of a creature's cognitive capacities should be used to mark the real difference between our cognition and that of spiders, insects, and worms[30] rather than the important, but comparatively superficial, differences between our cognition and that of our closest neighbors on the spectrum.

[30] However, Charles Darwin said of the mental qualities of worms,

> We have seen that worms are timid.... Judging by their eagerness for certain kinds of food, they must enjoy the pleasure of eating. Their sexual passion is strong enough to overcome for a time their dread of light. They perhaps have a trace of social feeling, for they are not disturbed by crawling over each other's bodies, and they sometimes lie in contact. (1911, p. 34).

Darwin also argued that worms "possess some degree of intelligence" (1911, p. 99). If one is inclined to be a Darwinian in this respect, one might wish to populate the far end of the spectrum with bacteria and algae instead.

Chapter Three

An Account of Theories Such That Children Might Have Them[1]

There has been a growing trend in developmental psychology to regard children as possessed of theories and to regard at least some of their cognitive development as similar to processes of theory change in science (Gopnik and Meltzoff 1997; Wellman 1990; Carey 1985; Perner 1991b; Kitcher 1988). Some proponents of this trend in developmental psychology have attempted to make clear exactly what they mean when they say of a child that she has a "theory," but they have found only limited help in the philosophy of science: The standard philosophical accounts of theories are not well-suited to the discussion of non-technical, everyday theories of the kind it is reasonable to think children might have. Psychologists have thus been forced into the position of developing their own accounts of what a theory is -- a useful and rewarding task, no doubt, but one matching more closely the job description of philosophers than psychologists. In this chapter, I will attempt to remedy this failure of philosophy of science to come to the aid of an actual science in need.

Specifically, I will offer an account of theories that -- unlike the accounts currently on offer in philosophy of science -- applies equally well to technically sophisticated scientific

[1] Parts of sections 3-4 have appeared in Schwitzgebel (1996), and are used here with the kind permission of *Philosophy of Science*.

theories and to the everyday theories of ordinary people. Only if we have an account of theories that applies to everyday theories will questions about the role of theories in the cognitive development of children be interesting questions with non-trivial answers. With such an account of theories in hand, I will spell out exactly the points of disagreement are between people who advocate the "theory theory" of development and those who do not. Finally, I will suggest a new domain of evidence by means of which to test the theory theory.

An account of theories broad enough to include within its scope both technical scientific theories and non-technical everyday theories also has value independently of any concern with developmental psychology. Philosophy of science can profit from an account of theories that reveals commonalities between scientific theories and everyday theories and thus captures some of the continuities between scientific practice and everyday life. Likewise, philosophy of mind can profit from a description of theories, to the extent theories play an important role in our cognitive lives.

In this chapter, then, I will present an account of theories that satisfies the following desiderata: (1.) It must make sense of the "theory theory" debate in developmental psychology: People who endorse the "theory theory" of development must hold that development crucially involves theories in my sense, and people who reject the theory theory must deny this involvement. (2.) The account must not lose sight of the fact that scientific theories are paradigm examples of theories, and it must

incorporate observations from philosophy of science into the discussion of theories. (3.) Good theories must in fact have most of the properties we take them to have -- they must be accurate, predictive, explanatory, revisable in light of new evidence, etc. (4.) The account must be clear and simple. In addition, I will claim for my account the following final virtue, not strictly necessary, but nonetheless useful for a variety of reasons: (5.) The extension of the term ‘theory’ on my account will map nicely into ordinary English usage. If, as I think, this fifth virtue holds, the account of theories I offer may be helpful as a starting point for other accounts of theories designed for other purposes.

1. The Axiomatic and Semantic Views of Theory

In recent years, the philosophical discussion about the nature of theories has taken the form of a debate between old-fashioned positivist views of theories (sometimes called the "axiomatic view of theories") and a newer approach developed by Suppes (1962, 1967), van Fraassen (1972, 1989b), Suppe (1977, 1989), Giere (1988), and others. The semantic view of theories is now in ascendancy within philosophy of science, although this ascendancy is not consistently recognized outside philosophy of science.

While I think great virtues may be claimed for the semantic view of theories, I will suggest that, in its substantive incarnations, it is too narrow to be a broadly useful account. Not only does it fail adequately to characterize non-scientific theories, but it applies awkwardly at best to many scientific theories as well (in developmental psychology, for example). Of special interest for my project, of course, is the question whether philosophical accounts of theories could possibly apply to the goings-on in the minds of young children. It would seem that neither the axiomatic nor the semantic views of theories, when construed substantively, could do so, since they both appear to require that those who subscribe to theories have a technical competence beyond that we can plausibly ascribe to young children.

According to the axiomatic view of scientific theories, a scientific theory has two parts. It consists first of a set of

axioms which, together with a mathematical and logical calculus, serve as the starting-point for the deduction of specific theoretical claims couched partly in theoretical vocabulary. Second, the theory contains a variety of "correspondence rules" or "bridge principles" relating the theoretical claims, which usually themselves cannot be directly tested, to directly testable claims couched entirely in logical and observational vocabulary. The function of a theory is to provide a basis for the deduction of particular empirically verifiable claims. These claims may come either in the form of predictions, if the truth of the claim has not yet been empirically verified, or explanations, if the truth of the claim is already known. (Explanation and prediction have the same logical form, the only difference being the evidential status of the deduced claim.) Proponents of the axiomatic view have differed with respect to some of the details of this picture, but the elements I have outlined were generally accepted by the central figures. Helpful expositions of the axiomatic view of theories can be found in Hempel (1952, 1965), Hempel and Oppenheim (1948), Carnap (1936/1954, 1966), Nagel (1979), and Suppe (1977, 1989).

Today, the "semantic" view of scientific theories, which I will describe in a moment, is more widely accepted than the axiomatic view. A variety of objections have served to repel philosophers from the axiomatic view, many of which are detailed in Suppe (1977). Among the more effective objections (to my mind) are:

(1.) The axiomatic view depends on a strict bifurcation of scientific vocabulary into “observational” and “theoretical” terms (the latter being partially interpreted in terms of the former by means of the correspondence rules). Even if one holds (for example, with van Fraassen 1980) that some clear sense can be made of an observable-theoretical distinction, it seems doubtful that this distinction can be made clearly in terms of a split in the *vocabulary* of science, as proponents of the axiomatic view have proposed. Consider the property of being round and the property of having an electric charge, the first apparently a clear example of an observable property, the second apparently a theoretical property. Nonetheless, there are cases of round things too small to be seen and for which, therefore, their roundness is not directly observable; likewise there are cases of electric charges sufficiently large to be directly observable, such as the charge I detect if I stick my finger in a light socket (Suppe 1989; Putnam 1962). Perhaps science could be given a new vocabulary that, in a non-circular way, divides itself properly between observational and theoretical terms, but such a project would be extremely complicated at best.

(2.) The attempt to provide an axiomatic, deductive system for even the most apparently axiomatic, deductive of sciences, theoretical physics, has generally met with only partial success, and the project has not been seen as particularly useful in the eyes of the scientists for whom it is supposed to be an aid (Suppe 1989; Cartwright 1983).

(3.) The account of explanation to which the axiomatic view is committed -- the view of explanations as deductions from laws covering the phenomena in question -- is plainly faulty. (For a painstakingly detailed history of the problems with this view of explanation, see Salmon 1989). It is possible both to have deduction from scientific laws without explanation (for example, one can deduce the height of a flagpole from the length of its shadow, the angle of the sun, and laws about the propogation of light, but one does not thereby *explain* the height of the flagpole) and to have explanation without deduction from scientific laws (consider the kinds of explanations that evolutionary biology provides: Evolutionary biology can often explain why a trait emerged in a population without necessarily having been able to deduce from prior laws that that trait would emerge).

The semantic view, in contrast, treats theories as *models*, or families of models, “isomorphic” to phenomena in the real world (or non-isomorphic if the theory fails). It is still, I think, a little difficult to discover exactly what a “model” is supposed to be on the semantic view (Downes 1993, for example, outlines some confusions), but at least in the most influential version of the semantic view, the “state space” view (elaborated in van Fraassen 1970, 1989b; Suppe 1989; Lloyd 1988), the interpretation is comparatively clear. In the state space version of the semantic view, a theory defines a system with some number N of *variables* that take a range of values (often numerical, but not

necessarily) and an N-dimensional *space* consisting of sets of ordered N-tuples of variable values. Each set of variable values is a logically possible state of the system. At any given time, the system will be in exactly one of its logically possible states, and the state it is in may change over time. The laws of the theory then serve to constrain either the evolution of possible states over time, or they may provide synchronic constraints on the set of states that a system may possibly occupy at a time. The ideal gas law (PV = nRT), for example, is a law of the latter sort, constraining the values that variable P can take given the values of V, n, and T (R is a constant). Newton's laws predicting changes in position for masses, given their velocities and accelerations, are laws of the former sort, constraining the change in values of the variables over time. Such laws may either be deterministic, like the ones I have cited, or probabilistic. When the model is used, some claim is made about structural similarities between the defined system and actual systems in the physical world. For the ideal gas law, for example, it could be claimed that if the physical system you are interested in modelling is an enclosed volume of gas, then the actual range of states it will occupy will, *ceteris paribus,* be a subset of the states allowable on the theoretical model, interpreting T as temperature in Kelvin, V as volume in cubic meters, and so forth. (Alternatively, one might wish to say that the actualy system will be *approximated* by a subset of allowable states, or *would be* in the subset of allowable states of the

system if the system were free from the influence of any but the indicated variables or parameters.) In such a case, one can say that the mathematical system described by the theory, or some substructure of it, is “isomorphic to” the physical system in question. It is also generally held that the physical data themselves to which the theory is applied are typically cleaned-up, idealized, and interpreted in the light of an understanding of the experiment from which they were obtained (Suppes 1962; Suppe 1989).

Quantitative theories in the sciences do, in fact, seem naturally suited to the semantic framework, and a number of people have attempted to show how evolutionary theory can be fit into the semantic model (Lloyd 1988; Thompson 1983; Beatty 1981). Evolutionary theory has been a particular focus in discussions of the semantic view, since it has seemed to some philosophers of biology particularly ill-suited to explication conformable with the axiomatic view of theories.

The semantic view of theories escapes the above-cited objections to the axiomatic view. It requires no strict distinction between observational and theoretical terms (although it is compatible with such a distinction); it does not require the axiomatization of scientific theories, and is compatible with -- even well-suited for -- current views regarding the idealized, *ceteris paribus* nature of scientific claims (Suppe 1989; Cartwright 1983); and it is not attached to the deductive view of prediction and explanation that has been so effectively

criticized since the heyday of positivism. Furthermore, it seems to do no violence to many scientific theories to characterize them as “models” in the above sense. Defenders of the semantic view have been fond of pointing out that state-space models look more like actual scientific systems than axiomatic systems do (Suppe 1989; van Fraassen 1989b; Lloyd 1988). We will see, however, that having as a desideratum that the philosophical explication of a theory look similar to the scientific presentation of it can also cut against the state-space view.

There are, nevertheless, a number of scientific theories -- especially theories whose primary weight does not rest on quantitative variables -- for which the semantic view does not seem particularly suitable. Consider, for example, Ellen Markman’s (1989) theory of lexical development in children. Markman notes that all children, in learning the meanings of words, must overcome “Quine’s problem” -- they must be able to learn, from relatively few encounters, exactly what class of things is supposed to be picked out by a single word. If an adult points to a rabbit and says “gavagai,” the child must determine whether the adult is referring to the rabbit, the rabbit’s ears, the color of the rabbit, the speed of the rabbit, the particular species of rabbit, the class of animals in general, or any of a number of the indefinitely many logical possibilities. Children are remarkably good at this daunting task and by the end of their second year are often able to guess the intended meaning of a word after a single use. How is this possible? Markman’s theory describes several tacit assumptions

children make about the meanings of words that serve dramatically to reduce the number of possibilities they must consider.

One important assumption children make according to Markman's theory is the assumption of "mutual exclusivity." The principle of mutual exclusivity demands that for each kind of object in the world, there be at most one label (parts of objects are thought of as distinct objects for these purposes, so that 'fin' and 'fish' do not stand in violation of the mutual exclusivity principle). Thus, for example, a child who hears a novel word will associate it with an object for which she does not already have a word, if one is present, rather than with an object for which she already has a word. Also, if an object with a known label is indicated by means of a novel word, the child will think that the word refers to something else, or to a part of the named object, rather than to the object itself. It follows from this principle that young children will have difficulty learning words that do not apply to "basic-level" categories (dog), but rather to superordinate or subordinate categories (animal, terrier), since to learn those words would require a violation of the mutual exclusivity assumption. The mutual exclusivity assumption would then work in conjunction with a variety of other assumptions to help constrain the range of meanings a child might judge a novel word to have.

Setting aside the question of whether Markman's theory of lexical acquisition is empirically well supported, we can ask how well it fits into the state-space semantic view of theories. I believe that this view can only awkwardly be made to fit. The

interesting parts of Markman's theories are not naturally thought of in terms of variables and constraints on variables, and do not seem to gain any clarity in being thought of that way: The theory is more prosaic than that. This is certainly not the way the theory is ordinarily conceived or described by its adherents and detractors. This latter point by itself is not necessarily an objection to understanding the theory that way: The positivists were happy to "explicate" a theory differently from the way praticing scientists understood it. Proponents of the semantic view of theories have not generally taken that stand -- they have held up the similarity between the scientific and the semantic understandings of their favorite theories as a virtue of the semantic account -- but there is no reason they couldn't take the positivist line in this matter. What would need to be shown in this case, then, is that the semantic view provides a better, more helpful understanding of theories like Markman's than the scientists' own understanding of it does. I suspect that this is unlikely, but I cannot of course anticipate every possible state-space approach to non-quantitative theories like Markman's, so I can only challenge the reader who is sympathetic to applying the state-space approach to such a theory to discover a useful state-space analysis of it.[2]

[2] For the curious reader, I have attempted to render Markman's theory into the language of the state-space semantic view. Here it goes:

> Let W be some unfamiliar word for the child in question, and let $\{O_1, O_2, \ldots O_n\}$ be the set of objects in the environment that are possible referents of W. Let $\{V_1, V_2, \ldots V_n\}$ be an index indicating, for each V_i the degree of preference for O_i as the referent of W, with the member of this set that takes the highest value being the assumed referent of W. Let $\{F_{i1}, F_{i2}, \ldots F_{im}\}$ take on values indicative of the presence or absence (or degree of presence) of various features of O_i relevant to its choice as the referent of W; for example, let $F_{i1} = 0$ if the O_i has no known name, and 1

The state-space version of the semantic view of theories seems even less applicable when we step outside science to everyday theories, such as conspiracy theories about J.F.K.'s assassination, Maxine's theory about why men are such jerks, implicit folk theories of psychology, physics, and so forth. The people holding such theories do not generally themselves conceive of their theories along the lines suggested by the state-space version of the semantic view. Many have no idea what a variable is, or a mathematical space, and probably some could not easily be taught to make sense of these ideas. The theories involved may not have clearly defined state variables or clearly defined ranges of value for their variables, and they may not be amenable to reconstruction in such terms without substantive change. I see no reason we should feel compelled to force such theories into the state-space mold, and I do not mean to suggest that advocates of the semantic view of theories would in fact suggest such a move. But then we are left with a choice between (a.) accepting the state-space view as a general account of theories and denying that everyday theories are in fact theories, and (b.)

---

if O_i has a known name. Markman's mutual exclusivity assumption can then be represented as the law: *ceteris paribus*, if $F_{i1} < F_{j1}$ then $V_i > V_j$.

If all of Markman's principles could be characterized in terms of relations between the F_{ij} and the V_i's, then Markman's theory could make do with only an (n*m + n) dimensional space (though I offer no promises here)! This seems an awfully complicated structure to saddle on Markman's simple theory. In addition, it offers some technical complications of its own. For example, what if the number of potential referents of W is a non-denumerable infinity (as seems likely)? Also, the account as stated suggests that the child (at least unconsciously) evaluates the plausibility of each object as a potential referent before making her choice, something not suggested by Markman's theory as originally presented. A state space account need not suggest that the child actually follows such a strategy: it could be revised so as to suggest that one of any number of non-exhaustive search strategies is performed by the child. What the state space account has more trouble accommodating is *silence* on the question as to the child's search strategy, a silence present in Markman's intended theory. An advocate of a state space interpretation of Markman could insist that although the theory is not silent as to search strategy, it merely "saves the phenomena" and is not intended to reflect the child's *actual* search strategy. This is inelegant: why introduce such unnecessary wheels?

rejecting the state-space view as a general account of theories. Such problems arise with double force for young children's theories, if young children do in fact have theories. If the "theory theory" of cognitive development treated theories along the lines suggested by the state-space view, I suspect it would have many fewer advocates than it does in fact have.

These objections are directed at the state-space version of the semantic view. Could perhaps another version of the semantic view weather such objections and make itself applicable to theories of all sorts, or at least scientific theories in general? Although many of the central exponents of the semantic view have spelled out the view in terms of state-spaces or other similarly mathematically, logically complicated structures, Giere (1988) has steered away from doing so.

As a consequence, however, it is not really clear what Giere means by "model" when he claims that scientific theories are families of models. He does not, in his general book on theories, models, and science, venture a definition of the term 'model' -- in fact, he says that he will be employing the term 'model' in more than one distinct sense (1988, p. 79). Some of the things he wants to call models are "abstract entities having all and only the properties ascribed to them," like the linear oscillators of physics (1988, p. 78). He also calls the contractionist picture of the formation of the Earth's crust a "model" (1988, p. 228). Elsewhere, he says that we make a theoretical model when we "imagine giving a party, including imagining who comes with whom and who says what to whom" (1989,

p. 27). Again, however, although Giere offers examples, he offers no definition. He does distinguish “theoretical models” of the sorts described from “analog” and “scale” models which are actually physical objects (1989, p. 23). However, one is left to wonder what all theoretical models have in common, besides their immateriality.

There may be some merit to Giere’s apparent evasiveness: Downes (1993) and Sloep and van der Steen (1987a&b) have argued that any substantive attempt to characterize precisely the formal structure of scientific theories will be apt to run across difficulties given the broad range of practices that seem to merit the title “scientific.” In particular, Downes argues, the claim that all scientific theories centrally involve models cannot reasonably be conjoined with any very specific idea of what a model is or what the relation between the model and the scientific practice is (or should be). The Markman example posed above points in that direction, as does Downes’ own example, the biological model of a cell (which looks even less mathematizable than Markman’s theory).

I am sympathetic with Downes’ suspicions. If the semantic view of theories is made sufficiently weak and deflationary, and if the notion of “model” is sufficiently broadened, then it may be true to say that all scientific theories involve models. If we want to go further and discuss not only scientific theories but also theories in general, scientific as well as ordinary, in all their different forms and sizes, we may well have to broaden the concept of a “model” so far as to grant (1.) that any set of

propositions defines a model (or familiy of models), and (2.) there is no kind of structure models necessarily have over and above the structure of the propositions that define them.[3] In this case, however, there would no longer seem to be much gained by invoking the idea of a "model." Why not just talk about the propositions instead?

In fact, this maximally deflationary semantic view has much in common with the view I will endorse below. But before getting to my positive account of theories, let's first turn our attention to what defenders of the theory theory have to say about the nature of theories. The views they defend, unsurprisingly, make it seem more plausible that children have theories than does either the axiomatic view or the state-space version of the semantic view. There is also a somewhat better match with the common-sense notion of what a theory is.

[3] Even this expansion won't be broad enough if we want to include actual physical models as "models" in the relevant sense, as suggested by Black (1962) and Griesemer (1990). The issue of the "structure" of propositions is a tricky one, and some views of that structure might undermine my point. For example, if a proposition has no structure beyond the set of "possible worlds" in which it is true, then all necessary propositions will have the same structure. Then, clearly, one might profit from using a structure of variables more fine-grained than propositions can be (for a dedicated attempt to reconciling our intuitions about the structure of propositions with a possible worlds approach to them, see Stalnaker 1984).

2. Developmental Accounts of Theories

It has recently become popular among developmental psychologists to characterize children, even very young children, as holding various “theories” about the world. Three- and four-year olds are said, for example, to be developing a “theory of mind” which helps them understand their own behavior and that of others (e.g., Flavell 1988; Wellman 1990; Perner 1991b). Likewise, a number of psychologists say that the conceptual changes involved with the development in the categorization of natural kinds are a result of “theory change” (e.g., Carey 1985; Gelman and Coley 1991; Keil 1991). A number of these psychologists have suggested that a useful analogy holds between the conditions and stages of theory change in science, as described, for example, by Thomas Kuhn (1962/1970), and the conditions and stages of theory change in the cognitive development of children (e.g., Gopnik 1988; Karmiloff-Smith 1988; Gopnik and Meltzoff 1997). Others (for example Spelke et al. 1992; Case and Okamoto 1996) have argued that the theory view of development is of limited application at best, and have proposed alternatives.

Naturally, it is useful in evaluating these claims to have a clear account of theories in mind. Ideally, one wants an account of theories that is neither so broad as to suggest that all mentation is theoretical, nor so narrow that only sophisticated academics can usefully be described as theoreticians. Unfortunately, the standard axiomatic and semantic accounts of

theories offered in the philosophy of science tend to fall into the latter camp, as should be evident from the discussion in the previous section. Surely no one but an academic could believe anything, for example, about isomorphisms to diachronic constraints on variables in an N-dimensional state space.

Developmental psychologists, then, have had to make do with home-spun accounts of what a theory is. I will now briefly sketch a few of these accounts and describe one of them in detail. I do so not merely for the purpose of canvassing the space of alternatives before presenting my own account, although this purpose might be sufficient in itself, but also because these accounts, I think, constitute a substantial original contribution to philosophy of science that should be appreciated in its own right.

Most of the developmental psychologists who have attempted to characterize theories have done so by describing two or more *features* commonly attributed to theories. It is often unclear whether these features are intended to constitute necessary conditions for something's being a theory, or sufficient conditions (taken jointly), or whether these features are to be seen as stereotypical characteristics of theories, in which case a thing is theory-like to the extent it satisfies the enumerated conditions. To the extent such questions about the developmental accounts are answerable at all, it is quite possible that some features are seen as necessary, some features as merely stereotypical, and some sets of features as jointly sufficient.

One simple characterization, in a paper instrumental in the recent burgeoning of the theory theory, may be found in a paper by David Premack and Guy Woodruff (1978):

> In saying that an individual has a theory of mind, we mean that the individual imputes mental states to himself and others.... A system of inferences of this kind is properly viewed as a theory, first, because such states are *not directly observable* and second, because the system can be used to make *predictions*, specifically, about the behavior of other organisms (p. 515, my itals.).

Here we see two conditions (apparently necessary conditions) on something's being a theory: (A.) It must refer to things that are not directly observable. (B.) It must be a system that can be used to make predictions. Adam Morton (1980) later expands the list of conditions to four, according to which a theory must (1.) aim to explain and predict phenomena, (2.) refer to individuals and properties lying behind the phenomena it is supposed to explain and predict, (3.) aim at the truth, and (4.) be open for public refutation. Similar accounts are given by Susan Carey (1985), Henry Wellman (1990), and Josef Perner (1991b), but the most detailed feature list, explained in the greatest depth, can be found in Alison Gopnik's and Andrew Meltzoff's (1997) work, to which I will now turn.

Gopnik and Meltzoff describe three classes of features characteristic of theories (1997, p. 34-41). They are:

Structural Features:

(S1.) Abstractness. Theories appeal to entities removed from or underlying the phenomena that provide the evidence for the theory. On their view of abstractness, gravity, planetary

orbits, and -- perhaps unintuitively -- bacteria and DNA all count as abstract.

(S2.) Coherence. Without specifying exactly what coherence is (a task which, admittedly, has proven tough for philosophers as well), Gopnik and Meltzoff suggest that theories exhibit some kind of internal coherence. As Morton says, if we were to number a wide variety of commonly held beliefs and examine the set of prime-numbered ones, they would likely not have the coherence essential to theories (1980, p. 6).

(S3.) Causality. Theories appeal to the causal structure thought to underlie regularities found in the phenomena in their domains.

(S4.) Counterfactuals. Theories support counterfactuals: They not only tell us what is the case, but they also tell us what would have been the case if....

(S5.) Ontological commitment. One is committed to believing in the real existence of the entities one invokes in one's theories.

Functional Features:

(F1.) Prediction. Theories generate predictions (or allow the people who hold them to generate predictions) about as yet undiscovered data in their domains.

(F2.) Interpretation. Theories allow their holders to interpret data and events in new ways. For example, advocates of one theory may consider the fluctuation of certain values as

crucial data to be accounted for, while advocates of another theory might treat those same fluctuations as mere noise.

(F3.) Explanation. A theory allows its holders to generate explanations of phenomena within its domain.

Dynamic Features:

(D1.) Denial. If someone holds a theory, a common initial reaction to (what an outsider might see as) counterevidence is denial. The potential counterevidence is ignored, or treated as noise, or treated as a problem to be worked out later.

(D2.) Ad Hoc Auxiliary Hypotheses. At a later stage, a proponent of the theory may attempt to rescue the theory from threatening anomalies by proposing ad hoc auxiliary hypotheses -- either adjustments and riders attached to the theory itself or claims about conditions in world surrounding the phenomena described by the theory.

(D3.) Alternative Models. Eventually, too many auxiliary hypotheses accumulate and the theory loses some of the simplicity and coherence that made it attractive in the first place, and people begin to consider alternative models of theories about the phenomena in question.

(D4.) Intense Experimentation and Observation. When (D3.) occurs, there is usually a period of intense experimentation and observation in attempt to adjudicate between the competing theories.

Although Gopnik's and Meltzoff's characterization of theories draws heavily from work in philosophy of science, it is

interestingly different from most of what has been done in that field. As far as I know, recent philosophers of science have either tried to characterize theories along roughly the axiomatic or semantic lines discussed above, or they have commented on individual features of theories or sets of features of the sort discussed by Gopnik and Meltzoff without explicitly attempting to address thereby the question of what a theory is, in general.

The feature-list approach to theories has some appeal, especially if one is attempting to capture the everyday notion of what a theory is. Our everyday notion, after all, seems likely to be a cluster concept of some sort, with candidates that possess a large proportion of theory-typical features counting as central examples of theories and candidates that have fewer of those features being more marginal examples. Nevertheless, 'theory' as it is used in the "theory theory" debate within developmental psychology, and in philosophy of science, is a technical term, and technical terms generally benefit from the clarity of being more precisely characterized than is typical for cluster concepts. (Consider the ordinary versus the scientific application of the term 'tree.')

If we look to Gopnik's and Meltzoff's list of features as a source of possible candidates for necessary features of theories, do we find anything that serves? Among those things that we would normally be inclined to call theories, we can find some that do not have one or another feature from Gopnik's and Meltzoff's list. Consider, for example, mathematical and philosophical theories. Although these certainly seem to be good

candidates for abstractness (S1.) and coherence (S2.), neither kind of theory generally appeals the causal structure (S3.) of events within its domain (2 + 2 does not cause 4). Other theories seem not to be abstract (my theory about why my car broke down), or to have little if any predictive power (F1.) (theories about the illnesses of the long dead or about why a certain battle was lost), or not to change in the way described above (D1.-D4.) (for example, if they are simply forgotten and replaced). Anti-realists in the philosophy of science (e.g., van Fraassen 1980) have argued against ontological commitment (S5.) as a necessary concomitant of theories.

If there are any plausible candidates for necessary conditions from Gopnik's and Meltzoff's list, they would seem to be (S2.) coherence, (S4.) counterfactuals, (F2.) interpretation, and (F3.) explanation. The first of these conditions is hard to deny, if hard to make precise. It does seem that every theory must have some degree of coherence, on any reasonable understanding of what coherence is. Likewise, it seems plausible to suppose that all theories support counterfactual claims (even mathematical theories: If this function had been such-and-such, the line would have crossed the x-axis here instead of there) as well as interpretions of some sort or other. Explanation, however, is of particular interest as a feature of theories. Many of the developmental discussions of theories have given it a central role (e.g., Carey 1985; Perner 1991b). Gopnik and Meltzoff also think that explanation has a special tie to theorizing:

> In fact, it may be that what we mean by saying that we've explained something is simply that we can give an abstract, coherent, causal account of it.... On the face of it, it would seem that one of the functions of a theory is to explain, and yet when we define explanation, we often seem to end up by simply saying that to explain something is to have a good theory of it, or to have some aspects of a good theory of it (1997, p. 38).

If what Gopnik and Meltzoff suggest is true, then explanation may not only be a necessary condition for having a theory, but it might come close, as no other feature seems to, to being a sufficient condition as well.

3. An Account of Theories

The main project of this chapter is to clarify the debate in developmental psychology over the legitimacy of saying that children have theories and that their cognitive development is a process of theory change. Toward this end, it is obviously useful to have a clear account of theories in hand, one that applies not only to sophisticated and technical theories in the sciences but also to the rough and ready theories of everyday life -- since certainly if children have theories, they must be theories of the latter sort. In the previous two sections we examined the accounts of theories on offer in philosophy of science and developmental psychology, and these accounts have been found less than ideal for the project at hand. The axiomatic view of theories that grew out of the positivist movement in philosophy of science fell to a series of objections widely known among contemporary philosophers of science. The semantic view of theories, the primary rival to the axiomatic view among philosophers of science, was found to be too narrow in its application, applying most helpfully to formal scientific theories containing mathematical variables, and not applying in any useful way to the informal theories of everyday life that are possibly to be found in children. The accounts of theories offered by developmental psychologists, most notably Gopnik and Meltzoff, consist primarily in feature lists, and although such accounts may accurately reflect our ordinary understanding of what it is to be a theory, I hope to present an account with

somewhat more precision and simplicity than the feature-list approaches have.

In this section, then, I will present a novel account of theories that I hope will adequately serve the project at hand. This account will connect theories closely with explanation. I will begin with a description of the account and a clarification of some of its features. I will then draw out some consequences of the account and in particular what is to be gained by the tight connection I postulate between theories and explanation.

The Account

This account will not be an account in the standard sense of a set of necessary and sufficient conditions, or even a list of prototypical features, but it is not for that reason any less valuable or any less specific an account. I will characterize what it is to *regard* something as a theory and what it is to *subscribe* to a theory.

(1.) A theory is a set of propositions.

(2.) Any set of propositions can potentially be regarded as a theory. To regard a set of propositions in this way is to be committed to evaluating that set of propositions in terms of its capacity to (allow subscribers to) generate good explanations in a domain.

(3.) To subscribe to a theory is to accept the propositions composing it and to employ them, or be disposed to

employ them, in explaining phenomena within the theory's domain.

Criterion (1.), that a theory must be a set of propositions, sounds more contentious than it is meant to be. I invoke the word 'proposition' on the understanding that I am using the word only in its "objects of belief" sense: I just want theories to be the kinds of things people can believe. In particular, I am not committed to seeing propositions either as really existing in some Platonic realm or as inherently linguistic entities. (So, for example, one might believe that Earl Grey tea tastes like *this*, or one might believe that riding a bicycle is done like *this*, without this knowledge being linguistically characterizable in any substantive way.) Furthermore, I think this account can be adapted at least to the semantic view of theories put forward by van Fraassen (1989b), Suppe (1989), and Giere (1988) and described above: The claim that such-and-such a family of models is isomorphic in the right way to such-and-such a range of phenomena (Giere's "theoretical hypothesis") is a proposition, if anything is. It is thus consistent with my account to agree with advocates of the semantic view about the crucial role models play in scientific theorizing, even if I cannot agree exactly with their ontology of theories. My focus is not on ontology, and my account can perhaps be adjusted to fit people's pet ontologies; (2.) and (3.), the conditions on regarding something as a theory and subscribing to a theory, are really the heart of my account.[4]

[4] I toyed with the idea that theories are in fact *logically (and mathematically) closed* sets of propositions because I didn't want it to be a result of my account that

Note that the three part account presented above only specifies one necessary condition for something's being a theory (that it be a set of propositions) and gives no sufficient conditions. Further specification of conditions would not, therefore, necessarily be hostile to my account. In discussing scientific theories, especially, one may be interested in adding further criteria. I am more interested, however, in what scientific and everyday theories have in common, and particularly in their psychological role. For the latter reason I focus on what it is to *regard* something as a theory and what it is to *subscribe* to a theory. I will now clarify a few things about conditions (2.) and (3.), which describe, respectively, these two aspects of the psychological role of theories.

Sets of propositions may be regarded as theories or, alternatively, as novels, or recipes, or laws, or editorial opinions. (For expository purposes, I am assuming leniency about inter- and intra-language translations.) Each of these classifications involves different criteria for evaluation. If I regard Marinetti's *Futurist's Cookbook* as a set of recipes to be

---

different but logically equivalent sets of propositions are different theories. Such a move, however, would have two counterintuitive consequences. First, no one would actually believe any theories. This difficulty could perhaps be finessed by the observation that people nonetheless often believe components of a theory from which the rest of the theory can be derived. Second, all theories true by virtue of their logical and mathematical properties alone would be equivalent (and would be components of every other theory as well). Appearances to the contrary, then, set theory and number theory would not truly be distinct theories, and no one would ever come up with a new, sound theory in mathematics or logic, but simply uncover new pieces of the One Theory. Similar problems would arise for self-contradictory theories.

Of course, if I do not require logical closure, my account is stuck with the consequence that logically equivalent theories are not identical theories, which seems a bit odd when one is a fairly obvious transformation of the other. Perhaps the best I can do to dispel this worry is to point out that people who believe obviously logically equivalent theories are each apt to believe the other's theory as well, and even if they don't, they are not apt to differ much in matters of substance within the scope of those theories. Thus, it is natural to be indifferent to which of two obviously logically equivalent theories is (for example) presented to a student, and to treat them as, for all practical purposes, the "same" theory.

evaluated in terms of the guidelines they offer for preparing good meals, I am apt to be disappointed. If I regard the very same work as a piece of modernist art, I might evaluate it quite differently. Propositions composing Orwell's *1984* might make very poor laws but excellent components of a novel and piece of social criticism. I am not prepared to describe sufficient conditions for something's being a law, recipe, or novel any more than I am ready to give them for something's being a theory, but it is immensely useful in understanding such things to explore the different criteria of evalution involved in regarding sets of propositions in any of these different ways.

This might seem a strange way of giving a philosophical account of a term -- discussing the criteria of evaluation one is committed to in applying that term to an object -- so I offer another example. Consider a body of water. If one regards that body of water as a fishing spot, one is committed to evaluating it in terms of its capacity to host a pleasant or productive fishing experience. If one regards that same body of water as a scuba diving site or a swimming hole, one will employ different criteria of evaluation. This is not to say that the *only* criteria by means of which one can evaluate a body of water regarded as a fishing spot are the criteria that make for good fishing spots -- one might, for instance, also think it would be a great place for a hydro-electric plant -- but one cannot ignore the fishing prospects in evaluating a body of water *qua* fishing spot. By understanding the different criteria of evaluation, we

understand just as well -- perhaps better -- what is meant when a body of water is referred to as a fishing spot or a dive site or a swimming hole than if we attempted to outline necessary and sufficient non-normative conditions for any of the above. Similar considerations apply, I think, to sets of propositions regarded as theories: They are better understood by outlining the criteria for their evaluation than by dwelling on what, precisely, should or should not count as an instance.

Finally, note that ordinary adults will, on this account, subscribe to theories about everyday things. Thus, suppose that Eric's car has broken down. He believes that it did so because the radiator was dirty and blocked, causing the coolant to overheat and the top radiator hose to blow, destroying all the belts and producing a shock that knocked loose the right front tie rod. Eric is disposed to explain a number of things about the breakdown and about the current state of his car by appeal to these facts, such as the loud exploding sound from under the hood immediately before the breakdown, followed several seconds later by a screeching sound and a strong pull to the right. Since he accepts the propositions described above and is disposed to employ them in explaining such facts, by criterion (3.) we can say that Eric has a theory about the breakdown. Similarly, Olga might have a theory about the assassination of J.F.K.: Oswald had co-conspirators within the government, he was set up to take the fall, etc., explaining the multiple bullet wounds, the failure of the investigation, and so forth. Unless an account of theories allows that ordinary adults should subscribe to such non-

technical theories, the debate over whether young children can have theories will be moot.

The Centrality of Explanation

On the proposed view, to regard a set of propositions as a theory is to be committed to evaluating those propositions in terms of what philosophers of science have called their "explanatory power." Good theories must provide good explanations. Bad theories, then, either provide bad explanations or no explanations at all. (The reader may decide for herself whether good explanations, and so good theories, must be true or approximately true.)

It might seem as though there are other evaluative dimensions besides explanatory power that I should be including in my account. After all, we evaluate theories not only in terms of their explanatory power, but also in terms of their beauty and simplicity, their ability to earn us grant money, and so forth. Still, I think there is something special about explanatory power that earns it the spot I give it in my account. In particular, I want to suggest that the demand that theories be explanatory can itself explain many of the other features commonly associated with good theories (turning van Fraassen 1980 on its head); that the linkage between theories and explanation accords well with ordinary usage; and that hooking theories to explanation in this way results in an account on which "subscribing to a theory" would seem to be an important kind of psychological state.

In discussing Gopnik and Meltzoff (1997) in the previous section, I granted plausibility to the claim that theories must be coherent, must support counterfactuals, and must provide their subscribers with the means to interpret events in the domain of the theory. I would now suggest that it is a mistake to regard these as necessary features of theories -- a *bad* theory, for example, might be incoherent or even self-contradictory in some way. Rather, what seems plausible is that *good* theories have all these features. Furthermore, all these features fall naturally out of the demand for explanation. Good explanations must appeal to some self-consistent, coherent base of facts. Good explanations allow those who understand them to understand and interpret the phenomena that have been explained. Good explanations provide a starting point for understanding not only what actually is the case, but also what would have been the case had some other conditions held.

Other features not strictly necessary for a theory to be good one, but nonetheless commonly associated with good theories, can be viewed as products of the demand for explanatory power. Good explanations often require appeal to the causal structure of phenomena; therefore, good theories often involve claims about that causal structure. When good explanations do not require appeal to causal structure, such as in mathematics, we find that the good theories in that area are not causal. Good theories tend to be predictive because, generally speaking, a theory would not be able to explain an event that occurred unless it could

have predicted it before it occurred (Hempel and Oppenheim 1948). And again, when explanation and prediction do fall apart (for examples, see Salmon 1989), we tend to associate theories with explanation. A non-explanatory predictive generalization (that Amir plays golf on Tuesdays and tennis on Wednesdays would in many contexts be such a generalization) is not ordinarily thought of as a good theory, while structures that explain but do not necessarily predict the events in their domains (such as parts of history, evolutionary theory, and psychodynamics) are often excellent theories. I suspect that many of the features we associate with theories -- if not all of them -- can be derived from the requirement that good theories provide good explanations. (These other virtues may also stand independently -- I do not require that theories *only* be evaluated in terms of their explanatory power.) The above account of theories, then, has the virtue of explaining a broad range of facts about the properties of good theories.

Explanation-Seeking Curiosity

I want to skirt as much as possible the raging debate in philosophy of science over the precise nature of explanation -- I think accounts of explanation that preserve most of our intuitions about instances of good explanation will also preserve the match between explantory power and theory quality. However, I do insist on one crucial feature of explanations: that they satisfy in us a certain kind of curiosity, what we might call an

“explanation-seeking” curiosity. (Some authors, such as Bromberger (1962), have even regarded this as a constitutive feature of explanations.)

If we grant that there is a kind of curiosity human beings have that is satisfied when an explanation is presented and understood, then it seems plausible to suppose that theories in the sense I am describing them play an important role in our mental lives. To *subscribe* to a theory is, I have suggested, to believe (or accept) the propositions of which the theory is composed and to be ready to use them in explaining phenomena in the theory’s domain. The curiosity that drives us to search for explanations will tend to emerge and re-emerge in a domain until we are capable of answering our own questions about that domain, i.e., until we subscribe to a theory that applies to that domain and can be used to generate explanations of the sort we seek. Explanation-seeking curiosity, then, will tend to drive us to the accumulation of (what we take to be) good theories; and to the extent this curiosity plays an important role in our mental lives, so also do theories.

I will now attempt to make this point a bit more precise, since it will play an important role in the application of my account of theories to the “theory theory” debate in developmental psychology.

The following conditions will serve to characterize a “drive.” An organism O has a drive toward goal G if O has the

tendency, from time to time, to enter a state of type S with the following features:

(i.) S leads O to engage in activities A_1, A_2, ... that in ordinary circumstances increase the likelihood of G;

(ii.) S has a characteristic subjective, phenomenal feel;

(iii.) there are characteristic circumstances C of which S is typically the product;

(iv.) at least some of A_1, A_2, ... are innate, unlearned responses to C;

(v.) G's achievement normally precipitates a (reinforcing) feeling of satisfaction, perhaps accompanied by a waning of S, especially if circumstances C no longer obtain.

The goal G will generally be a biological need, or a state or activity closely linked with a biological need. We have, for example, a drive to engage in sexual activity, closely linked with the need to reproduce; we have drives to eat and drive, closely linked to the needs for nutrition and water replenishment; drives to rest, to defacate, and so forth. These drives all meet the conditions described above: They have characteristic phenomenology and characteristic causes, they lead to activity increasing the likelihood of bringing about the goal, sometimes by innate, unlearned mechanisms, and the achievement of their ends brings a pleasant satisfaction. (A drive is unconscious if its characteristic phenomenology is not felt.)

Drives and desires are closely linked, but not identical. Typically, if a person is in the state S described above, that person desires the achievement of goal G. However, one can have

a drive toward a goal G even when one is not in S and does not desire G -- for example, a monk still has a drive to engage in sexual activity, even when he neither desires sexual activity nor (at the moment) feels the phenomenology characteristic of the sexual urge. Conversely, most adult human desires are not for anything that can be characterized as the goal of a drive. I might desire to bring my car in to get fixed on Thursday, but there are no characteristic circumstances of which this desire is typically a product, and there are no innate, unlearned responses that further the same goal. Furthermore, I would claim that such a desire has no characteristic phenomenology.[5] Perhaps the closest thing to a characteristic phenomenology would be the phenomenology of running a verbal image through one's head, something like, "boy I'd really like G." However, this seems hardly necessary, or even very common, for the possession of most desires, and certainly will not occur among creatures without language (and, of course, for people who only know languages besides English, such verbal images will have a different character). The relationship between such phenomenology and the desire to bring in one's car to get fixed on Thursday is nothing at all like the kind of relationship between the feeling of hunger and the drive to eat. It is really the latter kind of relationship that I regard as characteristic of drives.

Human beings have social and informational needs as well as immediate organic ones: It is our capacity to interact productively with each other and to acquire knowledge that gives

[5] For a similar argument regarding belief, see Chapter Two, p. ***.

us a reproductive edge. In response to these needs, evolution has imbued human beings with social and informational drives. The feeling of loneliness, for example, is associated with the drive to interact socially, and the feeling of curiosity is associated with the drive to acquire information about one's environment. It is important to notice in this regard that A_1, A_2, ... need not necessarily be *externally observable* activities: Private acts of cognition are just as legitimate a means to resolve curiosity as externally observable information-gathering.

Notice, also, that as human beings grow more socialized and sophisticated, their means of satisfying their drives and the phenomenology surrounding them will become more elaborate -- just look at the way a variety of social, informational, and biological drives get woven together in adult eating situations. This increasing sophistication does not, of course, mean that the original drives have been thrown overboard.

We can now give a little more substance to the claims with which I began this subsection. I wish to assert that people have a drive to seek the kind of knowledge conveyed by explanations, or, a little stronger, they have a drive to accumulate what they take to be good theories of the world around them. This drive produces exploratory behavior, hypothesis testing, question asking, and private cognitive activity of various sorts; it manifests itself phenomenally in explanation-seeking curiosity; and it is typically aroused when facts or patterns become salient

that the subject has difficulty assimilating into her present worldview.

A few remarks are in order about explanation-seeking curiosity as against other types of curiosity. Bromberger (1962) offers some good examples. I might, for instance, be curious just how tall Mt. Kilimanjaro is. In this case, I do not want an explanation of any sort: I want a number, in feet. On the other hand, if I am curious how water comes to emit bubbles as it heats in a pot, I want an explanation. Now are these really two different *kinds* of curiosity, and thus instances of two different informational drives? Or are they merely instances of the same phenomenological species, curiosity, only directed toward different objects?

I want to make it clear that my account does not hinge on one or another particular way of resolving these questions. Consider an analogy to hunger: Sometimes I am hungry for meat; sometimes I crave sweets. Are these two different kinds of hunger, two different drives, or one single drive directed toward two different kinds of object? To say that hunger for meat and craving for sweets are aspects of the same drive is to emphasize the similarities and the extent to which one kind of satisfaction might substitute for the other; to distinguish them is to emphasize their difference and non-interchangeability.

One more remark about explanation-seeking curiosity: Although explanations obviously satisfy this type of curiosity (hence the name), one need not always actually experience a linguistically

conveyed explanation for the curiosity to be resolved -- all that is required is that one acquires the type of understanding that would typically be conveyed in an explanatory episode.

If cognitive change is really theoretical in the fullest sense, then the drive to acquire knowledge that satisfies explanation-seeking curiosity must play an important role in the cognitive development of children. In the next section, I will argue that consideration of the affective and emotional consequences of the existence of such a drive in children should be considered an important source of evidence in evaluating the viability of the "theory theory" of cognitive development.

A Revision of (3.)

Before concluding this section, however, we should note one potential problem with (3.) above (that subscribing to a theory involves being disposed to employ the propositions of the theory in explaining phenomena within the theory's domain): It presupposes the capacity to convey what one understands in the form of an explanation. But this does not seem obviously necessary in order to subscribe to a theory. By the age three or four, children pretty plainly have explanation-seeking curiosity and can satisfy that curiosity by acquiring a broad understanding of the phenomena in question -- and so, I would like to say, they subscribe to theories -- even when they lack the capacity to explain the phenomena comprehensibly to an adult. One could also, I suppose, imagine examples of mute or painfully shy

creatures to whom we would wish to grant theoretical understanding without the capacity for explanation -- at least if explanation is regarded as a kind of linguistic act. (If explanations can be non-linguistic, internal actions directed toward the self, then perhaps these problems will not arise; but I do not want my account to depend on such a view of explanation.)

I would like, then, to alter the third element of the account of theories given above, at least as it applies to cases of the sort just described.

> (3′.) To subscribe to a theory is to accept the propositions composing it in such a way that acceptance of those propositions is causally sufficient, generally, to quell the pressure of explanation-seeking curiosity on the topic in question when facts explainable by the theory become salient.

I know this is an awkward mouthful, and because of its complications the original (3.) may serve as a more practicable criterion in standard situations. Let me explain a few of the clauses. Note that it may take a certain amount of time for the subject to realize that the salient facts are indeed explainable by the theory. Given the imperfection of our cognitive machinery, there will also certainly be cases in which the subject never realizes that the salient facts are explainable; thus, I have only required that explanation-seeking curiosity *generally* be mitigated. I have furthermore required that

acceptance of the propositions of the theory only be *causally sufficient* for the mitigation of curiosity to handle cases in which the explanation-seeking curiosity is not present for other reasons (such as being too hungry to find the topic worth thinking about), but *would* be mitigated by acceptance of the theory were the actually effective curiosity-stoppers not present.

4. Cognitive Development and Theories[6]

So, finally, should children be thought of as little scientists, whose cognitive development consists primarily in theory change, as suggested by, for example, Gopnik and Meltzoff (1997) and Henry Wellman (1990)? In this section, I will first describe a variety of developmental theories and the extent to which such theories can be said to treat cognitive development as theoretical. I will then suggest a new way of putting the theory theory to the test.

Some Views of Theories in Development

The debate over the "theory theory" has been marred by an inconsistent and variable understanding of what it is to subscribe to a theory, as well as a confusion among some of the proponents of the theory theory between three separate questions, namely, (a.) whether children subscribe to theories, (b.) whether the motor that moves cognitive development is the drive to revise and improve theories in the light of evidence that bears on them, and (c.) whether cognitive development consists primarily in domain-specific improvements in theories. Keeping these questions straight will help us in assessing the degree to which different theories of cognitive development treat development as theoretical. Note that the nature of development may differ from domain to domain. Language development, for example, may not be

[6] Much of this dissertation has been strongly influenced by Alison Gopnik, but this section even more than the rest grew from ideas planted in me by her.

at all theoretical, while the development of folk psychology may be theoretical in the fullest sense.

I will now examine a variety of approaches to development, with an eye to the three questions described above. To the extent that these questions are answered in the affirmative, I will regard the account as "theoretical." Some accounts will answer all three questions in the negative, and so make no appeal to theories at all; other accounts answer some of the questions in the affirmative, and so may be considered partially theoretical accounts of development. Those who endorse the theory theory in the fullest sense answer all three of the questions in the affirmative. I cannot here do full justice to the variety of accounts of development that have been offered, nor even to the subtleties of the accounts I do describe. My intention, rather, is to provide a rough idea of the spread of existing positions.

Let us begin with a sampling of accounts make no appeal to theories at all. So, for example, views that characterize development as the accumulation of particular empirical generalizations, or scripts (Shank and Abelson 1977; similarly, Nelson 1986), or narratives (Bruner 1992) make no appeal to theory-like structures. Take, for example, the idea of the script as it appears in Roger Shank and Robert Abelson (1977). Their classic example is the "restaurant script" -- essentially a set of generalizations about what precedes what in ordinary restaurants, providing the possessor of the script with a set of expectations allowing her to guide and interpret actions in a

restaurant context and to understand stories about restaurants. Shank and Abelson focus on the “coffee house track” of the restaurant script, which differs in details from, for example, the fast food track or the buffet track. The coffee house track of the restaurant script tells us that the first thing we do after entering a coffee house is scan for a vacant table in the smoking or non-smoking section (according to our wishes) and seat ourselves there. If there is no menu on the table, we can expect to be brought one promptly, and if this does not happen, we may flag down a waiter and request one. At such a time, we will probably be asked whether we would like anything to drink while we look over our menus and decide what we would like to eat... and so forth.

Such a script, although it offers predictions of what will happen in various circumstances, does not *explain* the events occuring in its domain: It tells us *that* they happen but not *why* they happen (Gopnik and Meltzoff 1997, p. 62-63). The restaurant script will tell us, for example, that we pay the owner of the restaurant rather than the owner paying us, but it will not tell us why this is the case. If someone is asked to explain why the owner gets our money, he will not (if he is truly interested in answering our question) merely appeal to the fact that this is what happens in the restaurant script; he will appeal to a *theory* -- i.e., a set of propositions to be evaluated in terms of their explanatory power. In this case, he would most likely appeal to a naive economic theory: In order to get the food, the owner has

to pay money to other people, so if she were to give it to us for free, she would be losing money, and that's no way to run a business. One does not really have a theory of restaurants until one can explain, and not merely list, the ordinary goings-on in restaurants. We can say, then, that scripts in this sense are *mere* empirical generalizations. To the extent development can be characterized as the acquisition of scripts, or script-like structures, it is not theoretical.

Simple connectionist models of development also probably should not be characterized as theoretical. A connectionist system consists of three or more layers of "nodes" which can take particular values and connections between the nodes that can take different "weights". A simple system will consist of a layer of input nodes, which are assigned different values as a way of representing some particular input; one or more "hidden layers" of nodes, whose values are determined as a function of the values of the nodes connected to them and the weights of those connections; and a layer of output nodes, whose values are determined as a function of the values of the hidden nodes and the connections weights leading from them to the output nodes, and whose values are interpreted as signifying some particular output or response to the input that was sent in. These connectionist networks are then "trained" by comparing the actual output with the desired output and modifying the connection weights in light of that output. (Paul Churchland (1990) has a helpful discussion of connectionism for beginners.)

A number of people have argued that development can usefully be modelled by connectionist networks (e.g., Bates and Elman 1993; Clark 1993; Karmiloff-Smith 1992). If connectionism is understood in a flat-footed way, it looks like development so characterized may require no appeal to theories. It certainly doesn't look on the face of it as though connectionist networks include theories, or representations, or beliefs. On the other hand, a more subtle view of connectionist networks may treat distributions of connection weights as somehow being representational or belief-like, and if this is the case, it at least opens up the possibility that connectionist networks can model aspects of development that look like theory-building (see, e.g., Bates et al. 1995). (This is not, of course, to say that connectionist networks *themselves* subscribe to theories, or have beliefs.)

The theory theory may also be contrasted with a modular or "central origins" view of development (Leslie 1994a&b; Spelke et al. 1992; Chomsky 1980), although the contrast is less stark. Spelke et al. describe the contrast in terms of the foundations from which cognitive development proceeds. On the central origins view, the primary source of knowledge in a domain is not sensory and motor experience but rather structures pre-existing in the mind from birth. Such structures may not be immediately available for use by the child, but only come "on line" as the child matures, perhaps as a result of triggers from the outside environment (Leslie 1988). These structures might even have a

variety of “parameters” that take one value or another, changing the nature of the application of the knowledge, depending on features of the environment -- Chomsky (1975) holds this position regarding grammatical knowledge. Such modules don’t have the all the kinds of characteristics that Gopnik and Meltzoff describe as central to theories: They are, for instance, innate and unrevisable and so lack the dynamic features of theories that capture their tendency to develop and change in the light of evidence. Gopnik and Meltzoff therefore conclude that modular knowledge is not theoretical (Gopnik and Meltzoff 1997, p. 56-59).

Nevertheless, some proponents of modular views want to describe modular knowledge as theoretical. Alan Leslie, for example, describes children as having a *Theory of Mind Module* (1988, 1994a&b). In his view, their knowledge is both modular *and* theoretical, despite the fact that it lacks the dynamic characteristics that Gopnik and Meltzoff regard as essential to theories. Here it is important to observe the difference between the three questions described at the beginning of this section: (a.) whether children subscribe to theories, (b.) whether the motor driving cognitive development is the drive to revise theories in the light of evidence, and (c.) whether development consists primarily in domain-specific improvements in theories. described at the beginning of this section. So long as the children can dispel their explanation-seeking curiosity about the mind by appeal to knowledge they have in the Theory of Mind

Module, they subscribe to a theory on the topic, by virtue of criterion (3′.) described in the previous section.[7] Thus, Alan Leslie can claim that knowledge of the mind is modular and innate, yet still be a theory-theorist in the weak sense of answering yes only to question (a.), the question whether children subscribe to theories. He cannot answer yes to either question (b.) or question (c.), however, since the modular view does not allow that evidence be a primary motor of cognitive development (b.) or that change in these theories is the meat of development (since the theories do not really change).

A modular view of development, then, is compatible with an attenuated version of the theory theory. Even so, it is unusual that the knowledge present in modules be accessible for the purpose of quenching explanation-seeking curiosity, as would be necessary for it to be theoretical knowledge on my account. So, for example, although many cognitive scientists believe that people have innate, modular, grammatical or visual knowledge, this knowledge is not available for explanatory use and so cannot, on the account I have presented, count as theoretical knowledge. People may act in some ways *as if* they had a theory about, for example, the necessity for anaphors to be bound by other expressions in their governing categories, but on my account we should not say that they *actually* have such theories. In chapter six I will present an account of belief on which it

[7] One might add the further condition that the knowledge be propositional; but in the extremely weak sense that I prefer to understand 'propositional,' *all* knowledge -- even know-how -- counts as propositional, since "propositions" are just whatever can be the contents of knowledge and belief. Furthermore, I see no reason why the things we know

will turn out, in fact, to be in some respects a dubious matter to ascribe this belief to (grammatically naive) people at all, independent of the question of whether the belief can be deployed to satisfying explanation-seeking curiosity.

We have seen that it is possible to answer no to questions (a.), (b.), and (c.), as those who hold script or narrative based accounts of development do. It is also possible to answer yes to (a.) but no to (b.) and (c.), as Leslie does. Jean Piaget (1952; Piaget and Inhelder 1969) provides an example of someone who says yes to question (a.) and (b.) but no to question (c.): Children, on his view (as I read it), are theoreticians driven by explanation-seeking curiosity to interact with and explore the world, and this interaction results in their cognitive development ((a.) and (b.)), but it does so by means of system-wide improvements in their cognitive abilities, rather than by domain-specific theory changes (c.). It is also, of course, possible to answer yes to all three of (a.), (b.), and (c.), as do Gopnik and Meltzoff (1997), Wellman (1990), and Carey (1985). Some of the predictions and expectations of such a view of development will be described in the next subsection. Of course, as noted above, it is possible to think that development in one domain is theoretical while development in other domains is not; when I say that Gopnik and Meltzoff, Wellman, and Carey endorse the strong version of the theory theory, then, I do not mean to imply that they do so for all areas of cognitive development.

---

when we have know-how, in other words these “propositions,” can’t figure in explanatory and theoretical activity.

Nor would, for example, Shank and Abelson necessarily endorse their script-based account of development as appropriate for all domains. Probably the most reasoned approach is a deliberate eclecticism.

A New Domain of Evidence for the Theory Theory

The full-blown theory theory of development, committed to all three of (a.), (b.), and (c.), makes the following reasonably well-publicized predictions about cognitive development (all compatible with the account of theories I offer):

(1.) Since theories are domain-specific, development should be domain-specific. For example, changes in one's theory of economic transactions should have only an indirect effect, at most, on one's biological theories, and we should not expect that transformations in the understanding of one domain will be synchronous with transformations in the understanding of other domains.

(2.) The pattern of development, in the domains to which the theory theory applies, should generally be from poorer theories (or no theories) to better theories, and the kinds of things leading to development should be the kinds of things leading to theory change, e.g., encounters with better theories or counterevidence that cannot easily be accommodated, as opposed to biological maturation or physical practice.

(3.) Cognitive structures in those domains should show the right degree of resistance to change. On the one hand, theories (unlike innate modules) are typically revisable, at least in principle, given enough clear counterevidence. On the other hand, people are naturally (and with good reason) reluctant to abandon powerful explanatory structures at the drop of a hat.

One problem with treating these three predictions as the core predictions of the full-blown theory theory, by means of which to distinguish it empirically from its competitors, is that the evidence adduced tends to be indecisive. Modular and script or narrative accounts also predict domain-specificity in development; all accounts of development predict increased understanding throughout childhood; and the generally negative results of attempts to induce broad cognitive change by offering counterevidence (except when the child is on the cusp of making the change anyway; see, e.g., Flavell et al. 1986; Resnick 1994; Vygotsky 1978) can be seen either as indicating innate modular constraints, or maturational unreadiness, or the natural reluctance to change theories given the limited amount of evidence an experimenter can present to a child.

Taking seriously the drive model I have suggested, I offer the following proposal that may provide a better means of empirically distinguishing the full-blown theory theory from its competitors: Look for the *patterns of affect and arousal* associated with the emergence and resolution of explanation-

seeking curiosity, and attempt to determine how the patterns relate to the cognitive development of the child. Let me offer an example of how this might work.

When a potential piece of counterevidence to a theory achieves salience, explanation-seeking curiosity will typically exert itself upon the child. The reaction might be characterized as something like a "why did that happen?" or "how is that possible?" reaction (though, of course, these words need not be uttered or internally produced). This reaction will typically be different, and often more prolonged, than the kind of surprise that follows a violation of expectations that offers no challenge to existing theoretical or explanation-producing capacities, such as the surprise one might feel at arriving home to find one's spouse has purchased a new toaster. It is also apt to produce a spurt of hypothesis formulation and testing, expressed either verbally or through physical experimentation. One might even, using Schacter's and Singer's (1962) or Zanna's and Cooper's (1974; also see Cooper and Fazio 1984) paradigm, attempt to determine whether curiosity-specific affect and behavior are reduced if the arousal can be attributed to some other feature of the environment.

If a new theory that accommodates the counterevidence is not developed, we may expect arousal to recur from time to time as the counterevidence presents itself again, even though the evidence itself may not be new, or even, any longer, unexpected. If the evidence is assimilated into the old theory or if a new theory is developed that accommodates the evidence, we might

expect a period of relief and/or excitement, resulting from the satisfaction of the explanation-seeking curiosity, and new instances of the counterevidence to the old theory should no longer bring arousal and curious affect. (And, of course, one would also expect the child to behave as though she believed the propositions composing the new theory.)

Such a pattern of affect, if it can be tied to the emergence and resolution of explanation-seeking curiosity, and if (1.) - (3.) above are also plausibly satisfied, would I think create a presumption in favor of the full-blown theory theory. Modular or associationistic views of development would not predict such a pattern of affect and arousal. This is not to say that people, especially as they grow older, might not have a diversity of reactions to counterevidence -- as I mentioned above, the instantiation and interweaving of drives can become complex -- but it would be an overreaction therefore to abandon the project of explaining patterns of action and affect by appeal to the drives behind them; as things get more complex, the project only becomes more difficult.

The theory theory has been successful in generating and making sense of much empirical research in cognitive development (a fact well demonstrated by Gopnik and Meltzoff 1997), but to the extent the battle has been fought primarily over the explanation of *cognitive* phenomena, the theory theory has missed a whole arena of potential support or disconfirmation in *affect*. If theories are psychologically real entities -- if children

really have them -- then they ought to find expression not only in cognitive patterns but also in patterns of affect. In deciding between theory-based and non-theory-based accounts of development, it would be a mistake to ignore this fact.

A final remark: If we grant that the same kind of curiosity driving this pattern of affect and behavior, and which I have called “explanation-seeking” curiosity, might be present even in primates or prelinguistic infants, then it may be possible to make some sense of the idea that even such creatures as these are theoreticians, seeking to satisfy their explanation-seeking curiosity by means of acquiring environmental information.

5. Conclusion

The primary aim of this chapter has been to develop an account of theories useful for addressing the "theory theory" debate in developmental psychology. In the first two sections, existing accounts of theories in philosophy of science and in developmental psychology were reviewed and found to be less than ideal for the goal at hand, for reasons summarized at the beginning of section three. A novel account of theories was then developed, centering around the questions of what it is to *regard* something as a theory and what it is to *subscribe* to a theory. The account proposed a tight connection between theories and explanation. In particular, it was argued that regarding a set of propositions as a theory commits one to evaluating those propositions in terms of their explanatory capabilities, and that to subscribe to a theory is to accept the propositions composing it and to be disposed to employ those propositions in satisfying explanation-seeking curiosity about the world around us. It was then argued that such explanation-seeking curiosity is what drives us to accumulate theories about the world. But if the accumulation of theories really is a product of such a drive, then that drive should manifest itself in patterns of affect and arousal associated with the development, testing, and refutation of theories -- and accounts of development that treat cognitive development as theory change ought to look for such patterns of affect and arousal. If such patterns cannot be found, then we should be hesitant to say that cognitive change really is theory-

driven in the way proponents of the full-blown theory theory suggest.

In the introduction, I set myself the goal of offering an account of theories useful both in clarifying the debate over the theory theory in developmental psychology and in furthering the goals of philosophy of science and philosophy of mind. Although the first goal was the primary focus of the chapter, I suggested that the second goal would be furthered by the development of an account of theories that captured some of the continuities between scientific practice and everyday life and that granted theories an important role in our cognitive lives. I believe that I have offered just such an account.

I would like to conclude by pointing out some implications of this account for the education of children. Science educators such as Hewson and Hewson (1984), di Sessa (1988), and Posner et al. (1982), while not always agreeing about the relative importance of theories in intuitive science, have generally agreed that *if* people have naive scientific theories, then the presentation of evidence conflicting with those theories ought to be of substantial use in leading them to acquire new, more accurate theories (at least to the extent that the conflict is recognized). The account at hand offers a mechanism by means of which such a process could work: Upon the presentation of the counterevidence, the student's explanation-seeking curiosity should be aroused, and she will be driven to construct a new theory, without which that curiosity could not reliably be

quenched. If the student's explanation-seeking curiosity is not aroused on the presentation of the counterevidence, then it may well be that she does not have anything as substantial as a theory about the topic in question, and the educator may wish to be directive in leading her to develop a theory. If explanation-seeking curiosity does arise, then perhaps the educator will benefit from employing the student's own drive to explain to generate interest and learning, with only the minimal guidance of a few well-chosen, intriguing examples or data points.

I have ventured no opinions about the means by which explanation-seeking curiosity can be induced in the absence of a theory with which data can conflict, but to the extent that the drive to explain is a powerful motivational force, educators would profit by discovering the means by which it can be cultivated, since, as I have argued, the most natural products of such a drive are evidence-sensitive, evolving, and improving theories. Once such theories are in place, they may have sufficient importance to the student even to lead to independent exploration and inquiry beyond the bounds of classroom assignments, should new challenges to those theories arise.

On the other hand, if the development and improvement of theories is typically the result of a drive to explain, certain perils for theory-development and learning also suggest themselves. So long as a person feels she has an adequate explanation of the salient phenomena, no explanation-seeking curiosity should be aroused, even if her theory is a weak one. Learning by the mechanism described is, therefore, hostage to

salience. Add to this the observation that people do not generally seem interested in searching for potential counterexamples to theories that are superficially adequate, and one has a recipe for stagnation. (It is interesting to note, however, that people do seem interested in the satisfaction they can get from discovering *confirming* instances that their theories explain (Nisbett and Ross 1980).) Furthermore, the drive to explain seems itself to be, for most people, a weak and tenuous drive compared with the drives to eat, to interact socially, to sleep, and so forth, and it is usually necessary that these other drives be sufficiently attended to before the drive to explain can get the play it needs to lead beyond rudimentary developmental accomplishments. The drive may also wane a bit as adulthood approaches -- whether by natural, internal processes or because of some environmental inhospitability -- unless it is actively and deliberately cultivated in the kind of relaxed, nurturing environment in which only a minority of people have the luxury to dwell.

Chapter Four

Representation and Desire: Case Study in How a Philosophical Error Can Have Consequences for Empirical Research

When Premack and Woodruff in 1978 asked whether the chimpanzee had a “theory of mind,” they prompted reactions not only from psychologists, but also from philosophers. Among those philosophers who responded to Premack and Woodruff were several who outlined a research paradigm for studying the understanding of false belief in primates and children (Bennett 1978; Dennett 1978; Harman 1978). This paradigm was later taken up by Wimmer and Perner (1983) and was instrumental in launching contemporary research on the child’s understanding of mental life.

Ever since, theory-of-mind research has shown how philosophical work can productively be employed by the practioners of other disciplines. There are risks, however; if the philosophy is genuinely being used, rather than merely tacked on as an afterthought, one would expect errors in philosophy to lead to further errors down the road. In this paper, I will examine one such error in theory-of-mind research, stemming from the misuse of the word ‘representation’.

What I shall argue, in particular, is the following. In contemporary philosophy, the word ‘representation’ is used with a variety of different meanings which are not always clearly distinguished even by the philosophers who discuss them. Some of

these meanings have found their way into the literature in developmental psychology, where they have been run together, resulting in equivocal arguments, misrepresentations of existing data, and even, I will assert, ill-fated research. I will begin by distinguishing two very different ways of viewing representation, and I will examine in detail how one philosopher conflated these different understandings. I will then describe the motivation and mistakes of the developmental research that is the focus of this paper. I will conclude with some suggestions about how certain experiments on the child's view of drawing might be of help confirming or disconfirming a popular hypothesis about the child's understanding of mind.

If this paper has any single effect on the reader, I hope it is this: That it entices her to acquire the (all too rare) habit of *clarifying* what is meant when the word 'representation' is employed, rather than simply invoking the word as though it had a single, univocal meaning on which everyone agreed.

Representation is a crucial concept in philosophy of mind and cognitive psychology, and trouble with its use is bound to strike to the roots of these disciplines. What I shall describe in this paper are only the troubles I know best.

1. Desire in Indicative and Contentive Accounts of Representation

The contemporary philosophers whose accounts of representation have had the most impact on the theory-of-mind literature in developmental psychology are probably Fred Dretske, John Searle, and Jerry Fodor. Although the differences between these philosophers' views of representation are enormous, this fact is not as widely recognized as it should be. (Even Fodor doesn't seem always to recognize the degree of difference between himself and Dretske; see Fodor 1984, 1987, 1990). I will focus on just one dimension of difference here, crucial yet typically ignored, and because ignored a source of unrecognized difficulties. The difference that interests me is the difference between *contentive* and *indicative* accounts of representation. Searle and Fodor offer contentive accounts, Dretske an indicative one.

I shall call an account of representation *contentive* just in case it treats as representational anything meeting the following condition:

(A.) It has propositional (alternatively: intentional or semantic) *content*.

The sense of 'propositional content' I mean to be invoking here is that now broadly used in philosophy of language and philosophy of mind. Although the notion of propositional content is notoriously unclear, my current project does not depend on any specific way of cashing out that concept. Accounts of the sort I want to label as 'contentive' are those that treat all the

following types of things as representational: beliefs, desires, and the other so-called “propositional attitudes”; sentences and linguistic acts; pictures, maps, and certain kinds of artistic objects perhaps. John Searle (1983), Jerry Fodor (1975, 1981, 1987, 1990, 1991), and Hartry Field (1978) offer -- at least to a first approximation -- contentive accounts of representation in the sense just described. Searle argues that anything with propositional content (everything listed above) is a representation. Fodor and Field argue that some things with propositional content, like beliefs and desires, while not themselves *representations* are nonetheless *representational states.* Belief and desires are “representational states,” on this view, because they are relations *between* people and internal representations. So, for example, John’s belief that it is raining is a relation between John and an internal representation with the content that it is raining (Fodor 1981, ch. 7; Field 1978).

Indicative accounts of representation require a further condition. Not only must any representation or representational state have “content” (condition (A.)), but also:

> (B.) The content of a representation *is supposed to match up* (alternatively, in “normal” conditions matches up) with the way things are in the world. If it does not, *misrepresentation* (itself a type of representation) has occurred.

On an indicative account, a representation's "job" is to reflect the way things stand in the world. All representations, on this view, have what Searle (1983) calls a "mind-to-world" or "word-to-world" (or "representation-to-world") *direction of fit.* This is in sharp contrast to things like desires and commands, which, though contentful, function *not* to reflect the way things are but (very roughly) the way things *should* be. Desires and commands have the opposite "direction of fit" -- they succeed by bringing the world into line with them, not by bringing themselves into line with the world. (For more on direction of fit see Searle 1983; Anscombe 1957; Humberstone 1992.) Fred Dretske (1988) espouses an indicative view of representation; so, for example, although he is happy to say that desires do have intentional content, he denies that they are representational (1988, p. 127).

Conditions (A.) and (B.) are meant to be approximate, not precise. Fodor, for example, though he accepts (A.) as a good "first approximation" of his view (1987, p. xi), suggests conditions in which he thinks having content is possible without representation (1987, p. 22). Searle seems to require that *mental* representations have not only a content but also a direction of fit (either direction), and a "psychological mode" (1983, p. 12). At the same time, Searle allows for "Intentional states" whose "representative content" is not a whole proposition. So, for example, Gernot might believe *that the stove is on* or desire *that Pauline arrive promptly,* but love *Sally* (1983, p. 6-7). Although belief and desire take entire

propositions as their contents, love does not. Since direction of fit is, for Searle, defined in terms of *propositional* content, Intentional states such as love, presumably, have no direction of fit, thus failing to fulfill one of Searle's apparent requirements for mental representations. Such details, however, are beside the point for my argument, so long as indicative and contentive accounts cluster *roughly* around the criteria I have given.

An essential point of agreement between those who subscribe to indicative and those who subscribe to contentive accounts of representation is that *beliefs* are representational. If I believe that yesterday it rained two inches, then I am in a mental state whose propositional content is that yesterday it rained two inches. If I believe that Rick will someday return my copy of Christopher Marlowe then I am in a state whose propositional content is that Rick will someday return my copy of Christopher Marlowe. Beliefs surely also satisfy condition (B.). My belief about yesterday's rain is supposed to reflect the way things actually are (or were) in the world. If it does not, it is my belief (not the world) that ought to be changed. Misrepresentation has occurred.

The crucial point of disagreement between the two accounts, for my purposes, is in the treatment of *desire*. On indicative views of representation (Dretske 1988, 1995; also Millikan's "indicative representation" 1984, 1993[1]) desire is not

[1] Millikan's distinction between "indicative" and "imperative" representations lines up nicely with my distinction between indicative and contentive accounts of representation

representational. Desires are not supposed to indicate how things are; they are dispositions or urges to bring things about that may not be the case, or hopes that events will transpire in one's favor regardless of one's input. We do not say of a person who desires an ice-cream sandwich but is not eating one that she *misrepresents* herself as eating an ice-cream sandwich. But if we regarded desires as representations in the indicative sense, we would be committed to saying that, by condition (B.) of representation: The content of that desire, "that I eat an ice-cream sandwich now," does not in fact match up with the world. Surely desires may be *based* upon false beliefs or misrepresentations -- perhaps I have forgotten what ice-cream sandwiches taste like and would be disappointed upon actually tasting one -- but that does not mean the desires themselves are misrepresentations. Rather, the beliefs that inform them are. Desires, then, are not representational states for those who subscribe to indicative accounts of representation. (For more on this, see Dretske 1988 and Millikan 1993.)

On *contentive* accounts, however, desires are clear-cut, central cases of representational states. Desires, like beliefs, are "propositional attitudes" *par excellence.* If I desire that Tori watch the sunset, then I am in a state whose content is the proposition that Tori watches the sunset. If I desire an ice-

(1993, p. 98-99). On indicative accounts of representation, only what Millikan would call indicative representations are representations. On contentive accounts, both her indicative and imperative representations are regarded as representational. Most of Millikan's discussions of representation are discussions of indicative representations.

cream sandwich, then I am in a state whose content is that I eat an ice-cream sandwich. You get the idea.

Both types of account draw on certain of our pretheoretical intuitions. Indicative accounts pull heavily on the idea that there are always things out in the world that representations are supposed to be representations *of* and that if those things are portrayed inaccurately, or if there are no such things to begin with, then the representation must be a *mis*-representation. Contentive accounts depend more on recognition of the possibility of fictional or hypothetical representations -- paintings, for example, that are "representations" of unicorns or military sandboxes that are "representations" of hypothetical manuevers. These ordinary-language intuitions about representation conflict with each other: One cannot grant full credit simultaneously to the idea that all representations are meant to be portrayals of the way things are and to the idea that representations can be fictional or hypothetical.[2] Hence the divergence between the accounts.

[2] An interesting intermediate case is representations of the way things *would be*. Such representations leave room for accuracy or inaccuracy of a sort, although they are not about the way things are. So, for example, one might misrepresent a unicorn as having a second horn, or one might make inaccurate claims about how the interview would have gone had you only not spilled your coffee. This would seem to be a fertile field for further exploration in the literature on representation.

2. An Example from Philosophy

Perhaps my exposition in the previous section of the distinction between indicative and contentive accounts of representation will seem obvious. Nevertheless, people do not always make clear when using the word ‘representation’ exactly what it is they have in mind. Philosophers of mind and, to an even greater extent, cognitive psychologists tend to use the word ‘representation’ unqualifiedly, as though everyone were in perfect accord over the meaning of that term. The term is far more frequently invoked than explained. Since the word has no univocal meaning in philosophy and cognitive science, such behavior is ill-advised. Not only are indicative and contentive accounts quite different in nature, but the contentive accounts are themselves quite different -- Fodor (1975), for example, thinks representations must have a formal syntactic structure, while Searle (1983) denies this. Add aesthetically-motivated accounts of representation (e.g., Wollheim 1993) and “re-presentation” puns (the latter sometimes offered by the very same authors who give different accounts of representation when the latter is not being contrasted with presentation; Searle 1983; Dennett 1991a), and you have a recipe for disaster. Shortly I will describe the errors in developmental psychology that are the focus of this paper. In this section I warm up with a similar confusion in Dennis Stampe’s article “Toward a Causal Theory of Linguistic Representation” (1977). This article had a substantial impact on later philosophical work on the topic of

representation (especially Dretske 1988, 1995; Millikan 1984, 1993; Fodor 1984, 1987, 1990), but to my knowledge no one has noticed Stampe's conflation of an indicative with a contentive understanding of representation.

Stampe's (1977) article ambitiously takes up the task of offering a "causal theory of representation," stated as generally as possible and intended to unify the then (and still) popular causal theories of knowledge, memory, belief, evidence, perception, and reference. What all these phenomena have in common, Stampe says, is that they involve a representational "object" (1977, p. 81). Understanding representation in general should then be of use in understanding these phenomena in particular.

Since Stampe talks about representations as being the kinds of things with "contents" and "objects" in a fairly traditional sense, it seems likely that he would be willing to accept something like condition (A.) on representation as described above. But is he also willing to accept (B.), thereby making his account an indicative one? Most of the phenomena mentioned on Stampe's p. 81 (cited above) could plausibly be interpreted as having a mind-to-world direction of fit (although the case of reference is not clear). If S knows that P, believes that P, has evidence that P, remembers or perceives that P, then S's mental contents are supposed to match up in the right kind of way with the world; if they don't, misrepresentation has occurred. If these are the phenomena in which Stampe is interested, then an indicative account of representation may be appropriate.

Stampe, however, hopes to include in his analysis not only the above-mentioned phenomena but also, it becomes clear as he proceeds, intentions and desires, as well as speech acts like promises and orders (1977, p. 82, 85). These latter phenomena have a world-to-mind (or world-to-word) direction of fit and, as discussed above, absolutely are *not* representations on indicative accounts.[3] To make his commitment to including such phenomena clear, Stampe says that the causal relation he wishes to make criterial for representation "is one that holds between a set of properties $F(f_1 \ldots f_n)$ of the thing (O) represented, and a set of propositions $\Phi(\phi_1 \ldots \phi_n)$ of the representation (R)"; and, he continues, the causal route may run in a number of directions and the relation still be a "representational" one (1977, p. 85). O's having F may cause R's having Φ, as in the case of true belief, or R's having Φ may cause O's having F, as in the case of an intention acted on and thereby satisfied, or there may be some common cause for both of them. It looks, then, as though Stampe's account might be a contentive one after all. He seems happy to ascribe representational status at least roughly to the same broad range of phenomena that Searle and Fodor do. (However, since Stampe does not explicitly say that he regards all items with propositional content as representations, we cannot be certain whether Stampe might wish to add some criterion that might exclude some, such as fears or doubts.)

[3] Stampe argues in later articles that desires do have an indicative function: The desire that P is supposed to indicate that *it would be good* if P were the case (Stampe 1986, 1987). Nevertheless, since the actual content of the desire, P, is not supposed to

Having said all this, Stampe remarks that “for the sake of having a manageable form of expression” he will “indiscriminately” just speak of the object as causing the representation (1977, p. 85). This is not at all an atypical move in philosophy and the cognitive sciences: We set ourselves the task of talking about “propositional attitudes” or “intentional states” in general (i.e., belief, desire, intention, fear, doubt, etc.); for simplicity’s sake we decide to talk about just one of them in depth; the one chosen “at random” is always belief; and we end up saying very little, except perhaps as a special study, about how the other propositional attitudes or intentional states are supposed to fit into our “general” account. It is particularly striking that we should see Stampe following this pattern, given the complex and detailed treatment of desire he develops in other work (Stampe 1986, 1987, 1994). But rather than focus on this later work of Stampe’s, which does not exhibit the tendency or error in which I am interested, I want to focus on the seminal and general 1977 paper of Stampe’s, since it displays quite clearly and usefully just the kind of slippage that proves damaging in the psychological work I will be examining shortly.

If Stampe wants to talk only of the object’s causing the representation, for the sake of having a manageable form of expression, but nevertheless wants his claims to apply to cases in which the causation runs in the other direction as well, then

---

match up with the world, even on Stampe’s account of desire, desires cannot be indicative representations as I have described them, by the criteria stated on p. 4.

it should always be possible to adjust his claims to fit these other cases. If his claims cannot be so adjusted, then he will have not done what he has advertised -- he will not have presented a *general* account of representation applicable to representations running in all directions of causation. One way of thinking about this potential error is as a conflation of indicative and contentive accounts of representation. The class of representations would be viewed widely, i.e., contentively, while the properties attributed to representations in general would include properties that apply only to indicative, belief-like representations in which features of the represented object cause a representation of that object as having those features.

Before Stampe even leaves page 85, he shows signs of having made the error in question. He says, for example, that "the causal criterion requires that the relevant properties of the object represented cause the instantiation of the relevant properties in the putative representation of it" (1977, p. 85). This *may* be a reasonable criterion to apply to belief, especially if one spruces it up with an account of misrepresentation (Stampe does so in terms of "normal" or "fidelity" conditions). There may be something funny about a belief that X is F that is not causally hooked up in the right kind of way with X's being F (although even Stampe wants to modify this claim when applied to beliefs about the future). But we cannot, as I have just argued we must, generate from this description an even remotely plausible analogous condition for desire. If X is not yet F, X's

being F cannot possibly cause my desire that X be F, since what does not exist cannot be a cause. Nor can we get the results we want by turning the direction of causation around. There is *nothing* odd or wrong about a desire that X be F that does not cause it to be the case that X is F. Some desires simply are not satisfied. Other desires, about the weather for example, we may hope to be satisfied, but not as a result of a causal chain involving the desire in question. Nor is it plausible to think that there must be some common cause of both the desire that X be F and its eventually being the case that X is F. Stampe's claim that "the causal criterion requires that the relevant properties of the object represented cause the relevant properties of the putative representation of it" would not seem plausible had Stampe "indiscriminately" chosen to talk of the representation causing the state of affairs represented rather than the other way around. Stampe already appears to have slipped into treating representation indicatively, attributing to all representations properties that do not rightly apply to representations contentively understood.[4]

From this point onward, Stampe's account looks like an indicative account of representation. On page 86, he says that "the central fact about representations" is that they "provide *information* about what they represent" (my ital.). But in what sense do, for example, promises that P, orders that P, or

[4] Stampe later argues that although what a desire that P represents is P, what it represents P *as* is a state of affairs the obtaining of which would be good (1987, p. 355). The desire is then "ideally caused by the fact that it would be good were P to be the case" (1986, p. 167). This is importantly different from the desire's being caused by P

intentions that P provide any information about the state of affairs they represent, P? On pages 87-90, Stampe has a discussion of what it means to say that a representation is *accurate*. It makes no sense to turn the causal direction of these ideas around and apply his discussions to non-indicative representations.

Furthermore, Stampe says:

> There is nothing essentially mentalistic about [representation]; it *may* be a wholly physical relation. Neither is there anything essentially *semantic* about it, in the narrower (proper) sense of the term. It is the relationship that obtains between the moon and its image reflected on the surface of a pond, and it would do so were no minds ever to have existed; even if there had been nothing to count them, the number of rings in the stump of a tree represent the age of the tree (1977, p. 87, his ital.).

If representation is disconnected from the mental like this and really can run either direction for Stampe, then it ought to be just as legitimate to turn things around and say that the moon represents the reflection in the pond and that the age of the tree represents the number of rings in its stump. Stampe, I assume, doesn't want to say this -- if he did say it, he would have to abandon the idea of any good match between his usage of 'representation' and anyone else's -- but there is nothing in Stampe's account of representation that suggests that the moon *can't* be the representation of the reflection. It seems doubtful that Stampe would have made an analogous claim had he chosen to speak consistently of the representation's causing the object represented rather than the other way around. Perhaps Stampe

---

itself, as would be required on Stampe's criterion cited in the text, which requires a

would want to add conditions to the representational relation meant to apply specifically to cases in which the causer is the representation, thereby ruling out cases like the moon's representing its reflection, but in fact he discusses no such conditions.

In sum, Stampe focuses in his 1977 paper on features of representation that apply to belief-like mind-to-world cases; as a result, his account of representation looks very much like an indicative one. This may be fine for most of the phenomena he wants to discuss in this paper, but he cannot apply his account to desire, intention, or any of a number of other phenomena with a world-to-representation direction of fit that he does in fact claim to cover with his account. Although the paper begins as if it were going to offer a contentive account of representation, the account looks more indicative in the end.

Stampe is not unique among philosophers in running together indicative and contentive approaches to representation, and I have chosen his 1977 article as a focus not to single out him in particular, but rather because it is an influential and clear example, and it shows how even a philosopher like Stampe, who is generally attuned to the complexities of desire and other mind-to-world representations, can slip into a belief bias when speaking broadly about representation in general. In the airy heights of abstraction and generalization, the difference between contentive and indicative accounts of representation can sometimes go unnoticed.

---

relation between P *itself* and the desire that P.

Even Fodor, whose remarks about representation are usually clearly in the contentive camp, sometimes slips into thinking of representation indicatively. The clearest case of this is probably in his 1984: On the first page, Fodor says that "the point about propositional attitudes [belief, desire, etc.] is that they are *representational* states" (1984, p. 231, his ital.) -- i.e., they are relations between people and internal mental representations (Fodor 1981, 1991). Fodor then, as usual, focuses most of his attention on representations with a mind-to-world direction of fit. Finally, on the closing page of the article, Fodor remarks that if R represents S, "what R represents is its truth condition, and its truth condition is whatever causes its tokening in teleologically normal situations" (1984, p. 249). With the indicative/contentive distinction in hand, we can see the difficulty here. The first quotation insists that desire is a representational state, but the second does not allow desires to involve normally tokened mental representations. When Fodor later rejects the position endorsed in the second quotation, he finds it necessary to spend an entire chapter *arguing* against the claim that "Normally caused intentional states ipso facto mean whatever causes them" (1990, p. 82, 89) -- an argument he surely would have found unnecessary had he reflected sufficiently on the fact that both he and those he takes to be his opponents regard *desires* as intentional states.

3. The Error in Theory of Mind

I would now like to suggest that a number of developmental psychologists studying the child's theory of mind have also conflated contentive and indicative approaches to representation. I will focus on the work of two of the most prominent (and most philosophically-minded) researchers in the field: Alison Gopnik and Josef Perner. I will begin with textual evidence that the word 'representation' is being used sometimes contentively and sometimes indicatively by these two authors. I will then show how equivocation between the two meanings of 'representation' produces problems for their research on the child's understanding of desire.

Lynd Forguson and Alison Gopnik begin their 1988 paper with a very clear statement of a contentive account of representation:

> Accordingly, we will understand by the term *mental representation* a mental state consisting of (a.) a *representational attitude* (e.g. believing, wanting, wishing, regretting, fearing), and (b.) a *symbolic content* ... that differentiates one belief from another, one desire from another, and so on (1988, p. 228, ital. theirs).

Notice that desire is specifically included in the list of representational states, since it has "symbolic content". Nonetheless, a few pages later Forguson and Gopnik say

> However, these children do not seem to be able to distinguish between the different informational relationships that may hold between representations and reality. As we will see, they show little understanding of the principles of representational change, representational diversity, or the appearance-reality distinction.
>
> All these abilities require that the child simultaneously consider a particular representation as a representation *and* as an indicator of how the world

really stands (1988, p. 234-235, ital. theirs, underlining mine).

These latter remarks only make sense on the view that all representations have an indicative function; one does not need to understand indicator relationships to understand that *desires* may change (Forguson's and Gopnik's "representational change") or that different people may have different *desires* (Forguson's and Gopnik's "representational diversity").[5] Forguson's and Gopnik's main thesis, in fact, depends on the slide between contentive and indicative accounts. On the basis of experiments suggesting a shift between ages three and four in the child's understanding of *indicative* relationships and misrepresentation, they argue that the four-year-old but not the three-year-old understands representation in general. This claim would be warranted if Forguson and Gopnik consistently held an indicative account of representation; it is not warranted if their account of representation is a contentive one. I will shortly describe in more detail the role this error plays in Gopnik's research, but first I will examine the work of one other researcher to make the point clear and to show the prevalence of the mistake.

Josef Perner (1991a&b) also seems to conflate contentive and indicative accounts of representation. He says, for example, that the "scientifically satisfactory" way to view a person's -- Sue's -- desiring something requires that "an internal representation is posited in Sue's mind, which represents the

[5] One might argue that desires *are* indicators of how the world really stands, a desire for food, for example, indicating a need for food (something like is Stampe's later (1986) view). Even if this were true, it's hard to see how it would be necessary to understand

nonexisting situation she desires" (1991b, p. 116) and that "treating desires as mental representations becomes necessary for understanding how desires change and how they are controlled" (1991b, p. 205). Thus, he sometimes seems to treat desires as clear cases of representations. Elsewhere, however, he says that "for any representation it is possible to misrepresent" (1991b, p. 20) and

> the definition of representation should therefore contain two elements: (a.) there must be a correspondence between states of the representational medium and states of the represented world, and (b.) this correspondence must be exploited by an interpretive system so that the representation is used as a stand-in for the represented (1991a, p. 144).

Neither of these latter remarks is consistent with regarding desires as representations: It makes no sense to talk of a desire as a misrepresentation of something (though the beliefs on which a desire *depends* may be misrepresentations); desires do not correspond the way beliefs do (or are supposed to) to states of the external, represented world; desires do not (in any clear sense) function as "stand-ins" for what they are supposed to represent.

Perner later argues (contra his 1991b, p. 205, cited above) that desires are not themselves *representations,* but rather are *representational states* consisting of relations between people and representations (1995; see also Fodor 1981 and Field 1978). On this view, S's desire that P is a relation between S and an internal representation whose content is P. This account of

---

this fact about desire to understand change and diversity in desires as Forguson and Gopnik suggest.

desires as representational states is also not consistent with Perner's indicative-sounding remarks about representation cited above (1991b, p. 20; 1991a, p. 144). My desire, for instance, that I get some fresh air is not plausibly seen as a relation between me and some internal mental thing corresponding with, and possibly misrepresenting, the state of the world. If it were, we would have to say that this desire of mine involves a misrepresentation: I am *not* getting fresh air, so any mental representation with the content that I get fresh air and the task of corresponding to the world would have to be failing in its representational task. But of course there is no misrepresentation. The facts are clear: I know that want fresh air, and I know that I am not getting it.

Perner, I think, recognizes that there is a problem here and seeks to escape it by arguing that desires involve a "secondary" type of representation:

> The primary function [of a representation] is to reflect the represented environment faithfully so that the user can learn to use it as a reliable guide. This is primary because it establishes the meaning of representational elements.... But once this meaning has been established, a map of a fictional environment can be generated by combining representational elements established by the primary process. This allows representations to be positively employed to represent hypothetical, nonexisting states of the environment (1991b, p. 24-25).

Perner follows these remarks with an interesting discussion of the use of "models" (e.g., a military sandbox) for both indicative and fictional purposes. However, although these remarks do clarify his position in some ways, they don't get him out of the bind described above: Either secondary representations

are truly representational, in which case his account is contentive and he ought not regard correspondence to the world and the possibility of misrepresentation as necessary attendants of representation; or secondary representations are not genuine representations, in which case desire ought to be left off the list of representational states.

The consequences of not deciding this issue are serious, since they lead Perner to some fundamental errors -- very nearly the same errors that Forguson and Gopnik make. Perner, like Forguson and Gopnik, sees the child as shifting, between ages three and four, from a nonrepresentational to a representational understanding of mind. (The title of his 1991 book, in fact, is *Understanding the Representational Mind.*) His argument for this depends entirely on evidence for a transformation in the child's understanding of facts *unique to indicative representations* -- i.e., that beliefs may be false, that appearances may differ from reality, that photographs may fail to capture the present situation. The conclusions Perner wants to draw, however, are supposed to apply to representations *contentively* understood, including desires and other mental states with a world-to-mind direction of fit.

Gopnik and Perner both have enormous influence on research in the child's understanding of mind, and so it is interesting to see them making such a similar mistake. But this mistake might be of merely conceptual interest, had it not also led to misguided empirical research.

It does so via the following equivocal argument, which both Gopnik and Perner accept:

(1.) Children come to understand representation at four years.

(2.) Therefore, their understanding of all representational states must undergo transformation at this age.

(3.) Desire is a representational state.

(4.) Therefore, the children's understanding of desire must undergo some important transformation (presumably analogous to their transformation in belief understanding) at four years.

First, some caveats. Neither Gopnik nor Perner put the argument forward in precisely this form. Nor does Gopnik, at least, deny the possibility of some "décalage" (difference in timing) between belief understanding and desire understanding (Astington and Gopnik 1991). They also each admit that there is probably some less sophisticated, "nonrepresentational" understanding of desire available to younger children. Gopnik sees no such nonrepresentational correlate for belief (Astington and Gopnik 1991); Perner argues for the existence of such a correlate, which he calls "prelief" (Perner, Baker, and Hutton 1994; Perner 1995). Nonetheless, in the final analysis Gopnik and Perner are both clearly committed to the equivocal argument just mentioned. They explicitly include desires in their lists of representational states, and they explicitly -- prominently -- declare that the child comes to understand representational states at four years. Unless desire is to be treated as a special case, more difficult

to understand than representational states as a whole -- a view neither Gopnik nor Perner endorse and against which there seems to be good developmental evidence (see below) -- the conclusion that desire understanding should change between ages three and four follows naturally.[6]

Now the problem with this argument is, as you may have gathered, that premise (1.) is warranted only on one understanding of representation, while premise (3.) is warranted

[6] Even some who do not buy into the dominant view described here may be committed to an analogous argument. Henry Wellman, for instance, (1990; Bartsch and Wellman 1995) similarly puts desire on his list of representational states and then ignores it in his more detailed discussions of representation. Since Wellman has studied the child's understanding of desire more extensively than most and has even given it a central role in his developmental account, this fact is especially surprising. In his most abstract discussions of representation, Wellman characterizes representations contentively, as states with "internal mental content" (Bartsch and Wellman 1995, p. 14). Wellman writes,

> In adult understanding as philosophers treat it, a person's desires are typically construed as similar to beliefs. Thus, both desires and beliefs are called propositional attitudes. Beliefs are beliefs about a proposition: Joe believes that that is an apple. In this construal, beliefs are understood as representational. "Joe believes that that is an apple" means something like that Joe has a cognitive representation of the world and in that representation the designated object is an apple. A person's desires can be construed similarly, that is, as desires about propositions, about possible represented states of affairs. "Joe wants an apple," then, is understood as something like, "Joe wants that there be an apple and that he obtain it." ... Since a person's desires are also representational in this sense, it is feasible to talk of desires for not-real, nonexistent imaginary things. We say things like "Joe wants a unicorn" or "Joe wants to be the best ski jumper ever" (Wellman 1990, p. 210).

Although Wellman also emphasizes a simplified, non-representational understanding of desire he thinks is available even to two-year-olds (Wellman 1990, p. 210-211; Bartsch and Wellman 1995, p. 13-14), he clearly thinks that the adult understanding of desire is fully representational: Desires are mental states taking full propositions as their contents.

On Wellman's view, the child comes to understand representation at around three years of age (in this, Wellman deviates from the majority view). One would thus expect the child's understanding of desires to become representational like the adult's, thus enabling the child to talk of desires for "not-real, nonexistent imaginary things." In discussing the transition from a non-representational to a representational understanding of mind, however, Wellman leaves desires out of the picture altogether. He repeatedly emphasizes that there are two sorts of representation: reality-oriented representations like beliefs and fictional representations like imaginings and dreams (Wellman 1990, ch. 9). Desires do not fit into either of these categories and are not mentioned. Thus, for example, in discussing the child's understanding of representational diversity, Wellman remarks that "even three-year-olds understand representational diversity, but they understand only the diversity allowed by imaginings and by a hit-or-miss understanding of misrepresentation" (Wellman 1990, p. 255). He says this in spite of the fact that he earlier presented studies (Wellman 1990, ch. 8) that, he argued, showed that the two- or three-year-old child could understand that people can have and act on desires different from the child's own. His discussion of the acquisition of an "active, interpretive understanding" of representation at four years of age similarly ignores desire: Although the child's understanding of false belief and the appearance-reality distinction are discussed at length, no attempt is made to examine the child's understanding of the active, interpretative dimensions of desire or even to discuss what such dimensions might be.

only on the other understanding. The argument is thus an equivocal one and invalid.

Gopnik's and Perner's arguments for (1.) depend on several experiments well-known in the theory of mind literature, and which have received broad attention in both psychology and philosophy. One classic is Gopnik's and Astington's "Smarties box" experiment (1988; also Perner, Leekam, and Wimmer 1987). Children are shown the easily recognizable opaque candy container for the English confection "Smarties" and are asked what they believe is in the container. Naturally the children answer "Smarties." The container is then opened to reveal not Smarties, but a pencil. Children are then asked a series of questions, including "When you first saw the box, before we opened it, what did you think was inside it?" and (in the Wimmer and Hartl 1991 version) "What will [your friend] say is in the box?" Three-year-old children, but not four-year-old children, typically respond "pencils" to both these questions.

Leaving aside the interesting methodological and theoretical issues this experiment raises, suffice to say that it, and others like it, are generally taken to suggest that the following competencies emerge at about four years of age: (a.) an appreciation that other people may have false beliefs (Wimmer and Perner 1983; Perner, Leekam, and Wimmer 1987; Moses and Flavell 1990); (b.) an appreciation that one's own beliefs may have been false in the past (Gopnik and Astington 1988; Wimmer and Hartl 1991); and (c.) an appreciation that things may appear to be other than they are (Flavell, Flavell, and Green 1983; Flavell,

Green, and Flavell 1986; Gopnik and Astington 1988; Friend and Davis 1993). That these developments should occur at roughly the same time is not surprising: They all seem to tap a basic understanding of the possibility of misrepresentation (but see Vinden 1996 for another view); and for many researchers, indeed, the child's coming to understand misrepresentation at that age is seen as the surest sign of her coming to understand representation then (Perner 1991b; Moses and Flavell 1990; Astington 1993; Olson 1988; but see Hala, Chandler, and Fritz 1991).

The important thing to notice here is that all these experiments tap abilities associated *exclusively with indicative, mind-to-world representations.* Desires cannot be false; desires cannot be misrepresentations. This kind of evidence, then, only warrants the first step of the argument described above if 'representation' is construed indicatively. But for step (3.) to be plausible, 'representation' must be understood contentively; hence, the equivocation. The same problem may be put another way: The experiments cited show (at best) that the child comes to understand the nature of misrepresentation at around age four; but this understanding has *no* bearing on the child's understanding of desire; the evidence so far supplied provides no reason to suppose that the child's understanding of desire ought to be transformed at this age. And in fact it is not.

Gopnik performed a number of experiments aimed at discovering the expected 3-4 shift in desire.[7] Astington, Gopnik, and O'Neill (1989; reported in Astington and Gopnik 1991), for example, performed an experiment to see if children were as poor at recalling their past unsatisfied desires as they seem to be at recalling their past beliefs. (Searle (1983) regards false beliefs and unsatisfied desires as structurally similar in that they both involve unmet "conditions of satisfaction.") Children were shown two toys that looked very different but could not be distinguished by touch, and asked which toy they preferred. The toys were then dropped together into a bag and the child was allowed to withdraw only one. The child was then asked whether she got the toy she had wanted. While almost 80% of three-year-olds correctly described their unsatisfied desires, they performed no better than chance on the standard (Gopnik and Astington 1988) test for recollection of past false beliefs.

One might object that there is no good way, in this experiment, to tell that the children aren't simply reporting on their *present* desire for the toy they didn't get. In the standard false belief recollection tasks, the belief is shown to be false and thus *changed* before the child is asked to recall it. The child sees the Smarties box, and it opened to reveal a pencil; the child's belief about the contents is thereby changed. The children are then asked what they had (falsely) thought was in the container before it was opened. In Astington, Gopnik, and

[7] That this was her goal is not only evident from the experiments themselves, but also has been confirmed by personal communication.

O'Neill (1989), on the other hand, the child's desire is not necessarily changed when the unwanted toy is withdrawn, and thus reporting their present dissatisfaction would be a successful response strategy. One might argue that it is this disanalogy, and not a fundamental difference in their level of understanding desire and belief, that explains the three-year-old's good performance on the desire task and poor performance on the belief task.

Perhaps with the idea of addressing this problem, Gopnik and Slaughter (1991) actually worked to induce a change of desire in children -- for example, by presenting them with two books, allowing them to choose one, and then reading it to them so that they then desired to hear the other book. They found that three-year-olds have some difficulty with reporting their past desires in this task, but not as much difficulty as with the false belief tasks. Notice, however, that this is no longer a test of their recollection of an *unsatisfied* desire, so again the parallel to false belief is not complete.

In another experiment, Gopnik and Seager (1988; again reported in Astington and Gopnik 1991) showed children two books, a child's book and an adult's book, and asked which book an adult would choose. A slender majority (57%) of three-year-olds claimed that the adult would choose the child's book. Four- and five-year-olds, on the other hand, said this only 36% and 28% of the time, respectively. Gopnik and Seager draw a parallel between these percentages and similar percentages one sees on the

false-belief tasks. They take the experiment as evidence that young children don't understand that different people can have different desires. This conclusion, however, is contravened by the results of other studies suggesting that children do have an understanding of the diversity of desires (Flavell, Flavell, Green, and Moses 1990; Repacholi and Gopnik 1996; Bartsch and Wellman 1995), and one wonders whether the results might be an artifact of children's not having a very good idea (or all *too* good an idea?) of what kinds of *books,* specifically, adults might care to read. It is interesting to see how hard it is to get the right kind of symmetry between a false-belief task, like the Smarties task, and any kind of desire task.[8]

Perner did not as actively (or at least not as publicly) engage in experiments directed toward finding a 3-4 shift in the child's understanding of desire. One experiment he did perform suggests that three-year-olds generally understand that people are happy when they get what they want and unhappy when they

[8] Moore, Jarrold, et al. (1995) similarly try to construct a desire task parallel to the false belief task. In their task, children are placed in competition with a toy character ("Fat Cat") to complete a three-piece puzzle. Both the child and the character begin the game with a puzzle piece for the body of a frog. Each needs to acquire, next, a head piece and, finally, the eyes. In order to win pieces, players must draw cards from a pack: a white card indicates that no action is to be taken, a red card indicates that one may take a head if a head is not already possessed, and a blue card indicates that one may take the eyes if one already has a head. The children and Fat Cat draw cards, and the child earns a head, but the puppet does not. Now, presumably, the child wants a blue card so that he may complete the puzzle. At this point, the child is asked two test questions: (1.) Which color card does Fat Cat want now? and (2.) Which color card did you want last time? These questions are intended to test that the child can understand both another person's desire that is different from his own and that his own previous different desires were different. Three-year-old children are found to pass this test in approximately the same proportions that they pass false belief tests.

This experiment is no more supportive of the thesis of a 3-4 shift in understanding the representational nature of desire than are Gopnik's experiments (and Moore et al. do not regard it as supporting this thesis). First, the parallel with false belief is not complete. These are not tests of *unsatisfied* desires, and perhaps are better compared to the child's understanding that people can have different beliefs when the facts of the matter are unknown, which seems to develop earlier than their understanding of false belief and to be in place by three years (Wellman 1990). Second, the task seems sufficiently complicated that it might introduce extraneous task-specific difficulties that could mask the three-year-old's ability to understand conflicting desires (an

don't (Hadwin and Perner 1991; see also Yuill 1984; Wellman and Banerjee 1991; Wellman and Bartsch 1988; Harris et al. 1989). In fact, the bulk of studies on the child's understanding of desire have found no important shift between ages three and four. Besides the studies cited so far suggesting that by age three children understand (a.) people's diversity of desires and (b.) their emotional reactions to the satisfaction or dissatisfaction of their desires, other studies suggest that three-year-olds also understand (c.) that desires can fail to match up with the world (Lillard and Flavell 1992) and (d.) that desires prompt action to obtain the object desired (Wellman 1990; Wellman and Bartsch 1988; Bartsch and Wellman 1989). That children understand desire substantially earlier than they understand belief is also suggested by their natural speech patterns (Bartsch and Wellman 1995; Bretherton and Beeghly 1982).

Probably because of his treatment of representation, however, Perner (1991b) seems committed to discovering a 3-4 shift in the child's understanding of desire. The best he can find is the Gopnik and Seager (1988) criticized above and a couple of experiments on understanding intention (Shultz, Wells, and Sarda 1980; Astington 1991; Astington 1993 makes a case that understanding intention ought to be regarded as of a piece with understanding desire). Astington's (1991, 1993) argument that the child's understanding of intention undergoes important changes at around the same time as her understanding of belief

---

understanding suggested by Flavell et al. 1990; Repacholi and Gopnik 1996; and Bartsch and Wellman 1995).

may in fact stand up to scrutiny. Moore, Gilbert, and Sapp (1995) also find something like a 3-4 shift in the child's ability correctly to distinguish "want" from "need". Of course, a skeptic might reply that it's not surprising that *something* changes in the child's understanding of such world-to-mind states around age four; what is more surprising, perhaps, is how *little* change there is.

I would like to end this section with some positive remarks about the current potential for productive interaction between philosophers and psychologists on the topic of representation and the child's theory of mind. A view of representation that seems to be quite popular in theory-of-mind research since the failure in the early 1990's to find a convincing 3-4 shift in the understanding of desire (*pace* Astington 1993) is neither a contentive *nor* an indicative one, but something somewhere in the middle, on which beliefs, photographs, maps, and other contentive items with a mind-to-world direction of fit are regarded as representations as well as (at least some among) images, fantasies, pretenses, and dreams, but *desires* are either explicitly excluded from the list of representations or conspicuously unmentioned (Leslie 1987, 1988, 1994a&b; Lillard and Flavell 1992; Olson and Campbell 1994; and sometimes, apparently, Wellman 1990). This approach to representation has yet to be justified or spelled out in any detail. A little philosophical work might be useful in making explicit what exactly the commitments of such a view are -- and whether there

is really a coherent, workable view here at all. Influence may run in the other direction as well. If it turns out that there are important developmental symmetries between understanding mind-to-world representations and some of these other representations -- symmetries that *don't* hold between either of these types of representation and desire -- then perhaps there is a useful category here that philosophers have missed and ought to begin to incorporate in their own work on understanding the human mind.

4. Representational Art as a Test of a Hypothesis About the Child's Understanding of Mind

Those who interpret 'representation' contentively have insufficient evidence to warrant the conclusion that children come to understand representation at age four, given the breadth of the class of representations the narrowness of the evidence base, as I have argued. But what if we read 'representation' indicatively? Should we see children as coming to understand *indicative* representations at age four? In this final section I will review some of the evidence for this conclusion, and I will suggest in rough outline an experiment that may help decide the issue.

As I have remarked already, the preponderance of developmental psychologists writing on the child's theory of mind see the child as coming to understand false belief and the appearance-reality distinction at age four, or possibly a little before. Various objections have been raised against this claim (e.g., Hala, Chandler, and Fritz 1991; Fodor 1992; Leslie 1994a&b; Lewis and Osborne 1990), but I will not attempt to assess their merit here. What I would like to focus on instead is whether, even accepting these experiments at face value, we have sufficient warrant to conclude that the child at age four comes to understand indicative representation *generally*. I think that the evidence is slender at best.

The first point to note is that the claim that the child comes to understand indicative representations at age four is *broader* than the claim that the child comes to understand the

indicative nature of belief at age four. More things than beliefs have indicative content. Popular candidates include assertions, maps, models, fuel gauges, drawings, and photographs, to name a few. If the child comes to understand indicative representation *in general* at age four, and not simply something about the capacity for *minds* (or eyes) to be mistaken or tricked (what the false-belief and appearance-reality tasks seem to test), we should expect some analogous transformation in the child's understanding of at least some of these other things at around four years of age. Although Judy DeLoache and Deborah Zaitchik have performed experiments that are sometimes viewed as a test of this hypothesis, I do not believe that the data warrant a conclusion one way or another about the timing of the child's understanding of indicative representation in non-mental domains.

Judy DeLoache's work on this topic (1989a&b, 1991, 1995) has primarily been on the child's understanding of models. In her classic experiment, she showed children a full size room with various items of furniture and a scale model of the room with miniature versions of the same furniture, arranged analogously, and she pointed out the correspondences to the children. She then introduced the children to "Big Snoopy" and "Little Snoopy" who liked to do the same things: If Big Snoopy was on the chair in the big room, Little Snoopy would be on the chair in the little room, and so forth. This correspondence was demonstrated for the children several times, and they were asked to place Little Snoopy in the appropriate place, given Big Snoopy's

location. The crucial test was this: The children were shown Little Snoopy hiding somewhere in the little room, and were told Big Snoopy would hide in the same place in the other room. The children were then instructed to find Big Snoopy (and were then requested to retrieve Little Snoopy as a memory control). If a child went directly to the analogous hiding place in the full-size room, she passed the test. If she searched randomly, she failed the test. Children were able to pass the task right around their third birthday. DeLoache's conclusion: They understand that the model (indicatively) "represents" or "stands for" the room (1989b, 1995). Since the children are only 36-38 months old, this is seen as an argument against viewing the 3-4 shift as a shift in the understanding of indicative representations.

Perner (1991b) has pointed out the flaw in this reasoning: Understanding correspondence is not equivalent to understanding representation. Note, for instance, that correspondence between A and B is a symmetrical relationship, while A's representing B is an asymmetrical relationship. Adapting an example of Perner's: In the tract-home suburbs of California, all the houses in a neighborhood are generally built according to one of four or five floor-plans. If I live in one such house, and I visit my neighbor whose house is built from the same floor-plans, I know exactly where the bathroom is. The houses, like DeLoache's models, correspond, but they certainly do not represent each other. Children, then, quite conceivably could understand the

correspondence between the room and model without understanding that one *indicatively represents* the other.[9]

Deborah Zaitchik's work (1990; see also Perner and Leekam 1990 reported in Perner 1991b) on the child's understanding of photographs is often cited as evidence for the generality of the child's transformation in representational understanding at age four. Zaitchik first familiarized children with a Polaroid camera, allowing them to take a picture and letting them watch the photo come out of the camera and develop. She then performed a skit with Sesame Street characters. She laid Ernie out on a mat in the sun and had Bert take a picture of him, which was turned face down and allowed to develop without the child seeing it. While the photo was developing, Big Bird came by and sat down on the mat. The children were then asked, "In the picture, who is lying on the mat?" Four-year-olds did well on this task; three-year-olds did not. Zaitchik argues that this experiment shows that the child comes to understand pictorial representations at the same time she comes to understand false beliefs -- and thus that we can characterize the child as coming

[9] DeLoache has argued against a "mere correspondence" interpretation of her research in DeLoache and Smith (forthcoming). DeLoache's and Smith's criticism of this view does not, I believe, succeed. First, it treats the mere correspondence interpretation as asserting that the children are only detecting simple correspondences between individual objects within the model and the full-size room. This, however, the view need not take this approach: Children might still understand the complex relation between the model room, its parts, and full-size room and its parts, even without understanding that the model symbolizes or represents the full-size room (again, consider the case of the tract-homes). Thus, DeLoache's arguments that children understand fairly complex relations between the model and the full-size room does not touch the question of whether they understand that one *represents* the other. DeLoache and Smith also assert that the correspondence view cannot handle later (but still similar) experiments of DeLoache's, but they do not describe why they think this is the case, and it is far from obvious to me.

to understand the nature of indicative representations in general at around four years of age.[10]

Other interpretations of Zaitchik's results suggest themselves, however. Understanding the operations of a Polaroid camera is neither necessary nor sufficient for understanding the nature of indicative representations. That it is not necessary is obvious: People who live in cultures without cameras will not understand Polaroid photos, but it would be wild to assume that they therefore do not understand indicative representation. The child has been given only the most rudimentary instruction in how this machine works. She might think that the picture will update to portray the current state of its subjects, or she might think that the picture portrays the way things were when it was developed, as opposed to when it was taken. Nor does knowledge of the working of cameras require the knowledge of indicative representation: The child can understand the correspondence between the photograph and the state of affairs at the time the picture was taken without understanding its *representational* nature, by an argument similar to the one presented against the DeLoache studies. If the child comes to understand Polaroids at about the same time she comes to understand false belief, I see no reason to suppose this to be anything more than a coincidence. In fact, Parkin and Perner (1997) find only very small and insignificant correlations between the performance of three- to

[10] Zaitchik, however, later argues that three-year-old children do have some tentative and wavering representational understanding of false belief (Zaitchik 1991).

five-year-olds on false belief tasks and their performance on a Zaitchik-like photo task.

Setting aside Zaitchik and DeLoache, then, the evidence for or against the claim that children come at age four to understand indicative representation generally, as opposed to indicative mental representations in particular, has been quite slender. A good test of this hypothesis is needed.

Some initial questions we might consider are: When does the child come to understand that models, or model toys, or very simple maps are *supposed* to match up with the things they represent and thus can be inaccurate?[11] When does the child understand that gauges and thermometers can misregister the properties they are supposed to detect? Dretske (1988) and Perner (1991b) have rightly emphasized the understanding of misrepresentation as the *sine qua non* of understanding the normative component of indicative representation. Unless the child understands the possibility of misrepresentation, one could argue that the child is simply picking up on the correspondence between the representer and the represented, not the essential fact that the representer is supposed to match up with the represented.

Lindsay Parkin and Josef Perner (1997) have recently performed some experiments testing the ability of children to understand misrepresentation outside the domain of the mental. In these experiments, children are tested on their ability to

understand that a *sign* (an arrow) might misrepresent reality, and their performance is compared with their performance on a standard false-belief task. So, for example, a story is told in which a train can either be at an engine house or in a tunnel. The child is introduced to a sign that is supposed to point to where the train is and a driver who has seen the train. The child then observes the train move from one location (where the sign indicates and the driver has seen) to the other (where the sign does not indicate and the driver has not seen). The child is then asked (a.) where the train really is and either (b.) where the sign shows the train to be or (c.) where the driver thinks the train is. The child who answers (a.) and (b.) correctly -- i.e. says that although the train is really in the tunnel, the sign shows the train as being at the engine house -- is scored as having understood the misrepresentational capacity of signs. The child who answers (a.) and (c.) correctly is scored as understanding that beliefs can be false. Parkin and Perner not only find a 3-4 shift in the child's understanding of misrepresentation in signs, but also find a high correlation between children's performance on the sign task and their performance on the standard false-belief task, even when age and their performance on a Zaitchik-like photo task are factored out. That the false sign and the false belief tasks should be found to be equally difficult is a little surprising, since the direction the sign indicates can be read right off the sign, whereas what

[11]Liben and Downs (1989) have studied child's understanding of representation in maps. They don't find any noteworthy understanding of maps before the school years, perhaps

the driver believes cannot be read right off any of his expressions. Still, perhaps this only shows how inattentive three-year-olds are to data suggesting the existence of misrepresentation -- something also dramatically brought out by Gopnik's and Astington's (1988) data suggesting that children will not report previous false beliefs, even if those beliefs were verbally expressed only moments before.[12]

Another place in which it seems natural to look for an understanding of misrepresentation, outside the domain of the mind, is in the child's understanding of the pictures she draws. The child's first drawings tend to be simple scribbles, but by age three or four, most children begin to produce what are commonly called "representational" drawings (Golomb 1992; Winner 1982; Arnheim 1974; Freeman 1980). These drawings, often of people, have distinguishable limbs and facial features, which are verbally labelled by the child as such. Although talk of "representation" is just as common among those discussing child art as among those discussing the child's understanding of mind, there has been little effort to connect these two fields and see

---

because of domain-specific task demands.

[12] Martin Doherty and Josef Perner (1997) also have recently found evidence that children come at four years to be able to monitor the use of synonyms, and that performance on this metalinguistic (and so arguably metarepresentational) task correlates with performance on the false belief task; but a test of the ability to monitor the use of synonyms is not a test of the capacity to *misrepresent* that is characteristic of indicative representations specifically, and so is less relevant to the argument of this section than the Parkin and Perner (1997) experiments. If Doherty's and Perner's data are interpreted as showing that children come at age four to understand representation, construed contentively, then the results will have to be reconciled with other experiments seeming to show an earlier understanding of desire. Alternatively, in accord with the suggestion with which I concluded section three, it may be that there is an understanding of representation that does not include desire but does include beliefs and a number of other things that are not specifically indicative, like words.

what light they might shed on each other, even by those whose interests cross the two areas.[13]

If it is right that an indicative understanding of representation comes to the child at age four, then a transformation in the child's understanding of her artwork ought to take place at around that time. It may be no accident that theory-of-mind researchers interested in child art have tended to push for earlier competence, perhaps in light of the three-year-old's "representational" approach to art (Sullivan and Winner 1991, 1993; Freeman, Lewis, and Doherty 1991; Freeman and Lacohée 1995), but they have not to my knowledge pursued the connection in any detail.

It is possible that the three-year-old or young four-year-old who shows little sign of understanding indicative representation according to the traditional tests may create "representational" drawings yet not understand their representational nature, i.e., the fact that, if one draws Daddy, some features of the drawing ought to correspond with features of Daddy -- if Daddy has two eyes the drawing ought not to have three, on pain of being a misrepresentation of him. To my knowledge, the child's understanding of this fact about drawings has not been systematically tested.[14] Anecdotal remarks suggest that at least five-year-olds understand that drawings can be "wrong" if they

[13] Notably, Ellen Winner (Winner 1992; Sullivan and Winner 1991) and Norman Freeman (Freeman 1980; Freeman 1991 makes some abstract and very general connections; Freeman and Lacohée 1995 uses photographs and pre-fab drawings as cues in false-belief tasks but doesn't use the child's own drawings or use misrepresentational drawings). Tony Charman (Charman and Baron-Cohen 1992, 1993) is an exception, but his research has primarily been on autistic children.

don't match up in the right way with the things they depict, and a view of early school-age children as determined to get their drawings "right" is assumed in some theories of artistic development (e.g., Willats 1984; Gardner and Wolf 1987). Golomb and Winner both provide examples (though they mean to draw something different from the passages here quoted than the child's understanding of the duty of the picture to match up with reality):

> James, age 5;4, draws a tadpole man with arms extending from the head. He looks at it attentively and remarks: "Never seen hands coming from the head" (Golomb 1992, p. 55).
>
> Conversation between an adult and a five-year-old:
> Adult: "Which is prettier, a flower or a picture of a flower?"
> Child: "A flower."
> Adult: "Always?"
> Child: "Yes."
> Adult: "Why?"
> Child: "Because artists sometimes mess up" (Winner 1982, p. 112).

It might be useful, then, to see at what age it is possible to elicit such remarks from a child, at what age they begin to criticize drawings that "get it wrong" about the objects they depict. Were we to find a 3-4 shift in this domain, that would, I think, provide dramatic confirmation of the claim that children come at age four to understand indicative representations generally. Failure to find an appropriate 3-4 shift, on the other hand, would suggest that the 3-4 transition is, at best, confined to the domain of indicative mental representations.

---

[14] Annette Karmiloff-Smith's (1990) study of children's facility at intentionally distorting their drawings is a start, but it does not specifically address the children's view of their own distortions.

A few potential pitfalls should be noted. First, there is what might be called the "Picasso problem." It is hardly straightforward business to discern when an artistic representation is a misrepresentation and when it is merely a simplification, a convention, or a creative distortion. If Picasso puts both of his subject's eyes on one side of her head, do we want necessarily to say that he is *misrepresenting* his subject as having both eyes on the same side? Similarly, if the child draws a "tadpole" figure with legs and arms proceding directly from what would appear to an adult to be the head, we may not want to leap immediately to the conclusion that this is a misrepresentation and hold the child at fault for not admitting this. Although adult "stick figures" look nothing at all like people, it is simplistic to say that they are misrepresentations.

A less obvious pitfall lies in the distinction between the child's noticing a lack of correspondence and the child's noticing a genuine misrepresentation. DeLoache's tasks, described above, suggest that the child understands that one thing may correspond to another from at least the age three (earlier with photographs: DeLoache 1991), but as I argued, this ought not be viewed as tantamount to understanding representation. One must therefore be careful to sort out mere observations of a lack of correspondence from genuine criticisms of a drawing as misrepresentational. (The Golomb quote above, in fact, is ambiguous in this way.)

Yet another pitfall is suggested by the second quote above: Deviation from intention or from convention may be seen as “messing up” -- e.g. if a line goes off the page -- without being understood as misrepresentational. It therefore needs to be made clear exactly why the child criticizes any particular drawing. If the child criticizes a drawing of Daddy with three eyes, is this because the drawing doesn’t correspond as it should to Daddy’s features, or is it simply that a certain convention -- two eyes per head -- has been violated?

Avoiding all these pitfalls in coming to understand the child’s view of drawing would be no trivial task, but the rewards in understanding how the child thinks would, I believe, be enormous.

5. Conclusion

In this paper I argued that philosophical accounts of representation could be divided into two rough camps: broad or ‘contentive’ accounts on which desire is regarded as a representational state (Searle, Fodor) and narrow or ‘indicative’ accounts on which it is not (Dretske). These accounts have not always been clearly distinguished, even by philosophers instrumental in their development (Stampe, Fodor). I argued that influential researchers studying the child’s “theory of mind” (Gopnik, Perner) have conflated these two accounts and, as a result, have been lured into misguided research on the nature of desire. I concluded with a positive suggestion on how research on the child’s understanding of art might confirm or disconfirm a popular explanation of the apparent shift between ages three and four in the child’s theory of mind.

Chapter Five

Toward a Developmental Account of Belief

An infant does not emerge from the womb knowing that winter is colder than summer. Yet by the time the child is eight, she believes this. One can imagine this belief in some cases coming to the child all in an instant: She has noticed that it is much colder these days than it was a few months ago; she asks why; she receives a full discourse on what it is to be a season, what winter and summer are, and that winter months are colder than summer ones (in non-equatorial climates). Suddenly, something clicks and she has the belief. But this is not the normal case. Knowledge of the seasons, like much of the child's knowledge, is more often acquired gradually. The necessary competencies and concepts are slowly developed. Bits of evidence are collected and falteringly put together. At the beginning of the process, we can straightforwardly say the child does not have the belief; at the end, she does have it. But in the middle, in the hurly-burly of development, it is neither wholly correct to say that she has the belief, nor wholly correct to say that she does not.

Epistemologists and philosophers of mind interested in belief have typically attended to the *instantaneous* (or nearly instantaneous) acquisition of beliefs as a result of the ordinary processes of perception and reasoning in adults. Rarely have philosophers attended to the more gradual processes of belief

development evident especially in young children. But surely it is not only children who experience the gradual development of beliefs: A college student might gradually come to believe that all the best speculative metaphysicians lived before the twentieth century, this belief growing slowly apace the student's understanding of what metaphysics is and her knowledge of philosophical literature. Before taking any courses in philosophy, our student had no beliefs whatsoever on the question of when the best speculative metaphysicians lived; it even seems misleading to say, as some Bayesians might, that she believed to some low or intermediate degree that all the best speculative metaphysicians lived before the twentieth century, and that her degree of belief in this proposition gradually increased with her philosophical education. It seems more accurate to say that before her philosophical education she had no beliefs at all, of any degree of certainty, about the pinnacles of speculative metaphysics; that by the time she graduated she did believe that the best speculative metaphysicians lived before the twentieth century; and that there was no single moment at which this belief established itself in her mind.

One of the great advantages of examining philosophy of mind through the lens of developmental psychology is that it forces us to recognize the importance of such *in-between* states of believing, states in which it is neither wholly accurate to describe the subject as believing the proposition in question, nor wholly accurate to describe her as not believing it. Such

states are, I would suggest, quite common in the gradual development of a new view, a new theory, or a new set of conceptual tools. When a person is in such an in-between state regarding some proposition *P,* the question "Does she believe that *P* or not?" plausibly cannot be answered with a simple yes or no.

Developmental psychology turns our attention to such states and demands an account of belief that takes such states seriously. Nevertheless, it would be a mistake to assume that in-between cases of belief are limited to situations of gradual belief development. The coming three chapters will all cover the topic of belief with a special eye to in-between cases of believing. As we proceed, I hope it will become evident that cases such as those of self-deception, of unconscious belief, and of belief poorly thought through can provide us with many examples of in-between believing.

What we need, and what philosophers have yet to provide, is a workable account of belief that presents a framework for understanding and classifying these in-between states of believing. In the chapter following this one, I will offer such an account. In the present chapter, I will lay some of the groundwork for that account. I will outline desiderata for the account, and I will warn against a class of intuitions and metaphors that run opposite the developmental and the in-between in belief.

1. Aims of the Account

I propose, as I have said, to offer an account of belief. Let us now clarify what exactly it is I take myself to be doing and what the criteria for success in my project will be.

Accounts are sometimes said to be given of *terms,* sometimes of *concepts,* and sometimes of *things.* Philosophers have not always been as careful as they might be in distinguishing the various different projects suggested by describing the analysandum in these different ways. It is one thing to give an account of the word ‘belief’, another thing to give an account of the concept of belief, and yet another to give an account of beliefs themselves. The first is a linguistic inquiry into the word ‘belief’, the second an inquiry into how some class of people think about belief, while the last is an ontological inquiry into the nature of belief. While one might argue that there are important relations between these three projects, it is hardly plausible to regard them as identical.[1]

My project in these chapters on belief has elements of each of the three dimensions described. Linguistically and conceptually what I am offering is a *recommendation*. I am suggesting that (English-speaking) philosophers and psychologists take up the habit of using the word ‘belief’ in the way I recommend and that they modify their concept of belief to match with the concept described below. It is *not* my project to provide an analysis of what we ordinarily mean by the word

'belief' or how, intuitively, we think of it. Despite this, one can hardly avoid talk of intuitions, and, for reasons I will soon mention, my account matches fairly well with ordinary, pre-philosophical intuition and usage.

My account is ontological to the extent it makes claims about the real world, as opposed simply to treating our way of thinking and talking about the world. I shall, for example, argue that there is no fact of the matter beyond a person's dispositional make-up about what that person really believes. I shall also argue for the pervasiveness of cases of in-between believing of the type alluded to in the introduction to this chapter. The first of these ontological claims will probably be seen as metaphysical, and I have no objection to so regarding it; the second claim is clearly an empirical one. I will not attempt to keep metaphysical and empirical claims separate, but will rather weave them together into my picture of belief. Indeed, it may be that the metaphysical and empirical shade into or cross-cut each other and that their separation would be ill-advised in any case.

The conceptual and the ontological elements of this account are supposed to support each other. It is because I think that certain facts about the world obtain that I recommend a certain concept of belief, yet it may be difficult to see that those facts obtain or to describe them without antecedently accepting the recommended concept of belief. This is not circular. It is not that the account depends on the truth of claims whose truth

[1] Discussion of the nature of analysis and the relation of language, concepts, and ontology was once more lively and sophisticated than it now is; for a useful historical

in turn depends on the truth of the account; rather, the merit of the conceptual recommendations of the account depends on the truth of ontological claims whose truth it may be *difficult to see* before accepting the conceptual recommendations. In part this is because the ontological claims one can make or understand depend on the concepts and words available. In part it is because one's regular habits of thinking greatly influence how one sees and structures one's experience of the world, even when new tools are made available. The reader may notice such an intertwining of conceptual and ontological issues in my treatment of in-between cases of believing: The attractiveness of my dispositional conceptualization of belief depends on the importance and pervasiveness of in-between cases, but someone who begins with a non-dispositional, all-or-nothing picture of belief may have trouble envisioning many of the cases described as genuine in-between cases. I hope to remedy this problem with a thorough attack on the all-or-nothing view and a plethora of examples.

If these are the elements of my account, what should count as success? I am not, as I have said, offering the account as an analysis of our ordinary concept of belief, so the primary standard against which the account should be gauged is not its match with ordinary intuition. Since the account is offered as a candidate for a *novel* way to think about belief, the criteria for success must be appropriate to this different purpose. First, I would hope that those claims that can be evaluated for truth or

---

account, see Urmson (1956).

falsity -- that is, primarily, the ontological claims -- are, in the main, true, or at least warranted, justifiable, and empirically adequate. Just as important, however, are the conceptual and linguistic recommendations of the account, which like all recommendations are not so much true or false as helpful or unhelpful. To count as successful, these recommendations must engender, or at least be apt to engender, good philosophical and scientific research. Something like this latter criterion, I think, should be a standard of success for *any* account with a stipulative dimension -- or, indeed, for any ordinary language account to the extent that the account is meant to be employed productively by philosophers and scientists, rather than simply marvelled at as a feat of linguistic analysis. As always, I will pay particular attention to the utility and practicality of the account for developmental psychology. I will argue, in particular, that the account will excel in its treatment of in-between cases of believing, which are prevalent in developmental psychology and which most standard accounts of belief are ill-equipped to handle.

A time may come when science and philosophy need not advert to such folksy things as beliefs in explaining mental life and behavior, as Churchland (1981) and Stich (1983) have suggested. If this is the case, then when that time comes accounts of belief of the sort I offer will serve no important scientific or philosophical purpose, unless it be merely to understand how deeply confused ordinary folk have been about the mind. If the

time for the rejection of folk psychology is now at hand, then the enterprise I have described is misguided: Science and philosophy will *not* profit from a new account of belief, and may even perhaps be hindered by it, as Marx felt the proletarian cause was hindered by the kinds of temporary capitalist palliatives that served to postpone the coming revolution. Better to let the concept alone, that we may sooner be inclined to cast it aside in favor of the new language of cognitive science.

While I do not think such a revolution is impossible, I fear it must be a long way off, if ever it will come. Although psychological and neurological research has overturned folk psychology at the fringes and in some narrow domains, scientists have so far not even come close to providing an alternative vocabulary with the broad utility that belief and desire talk has in folk psychology. Folk psychology is, in truth, a sophisticated, long-tested, highly accurate, and evolving theory, and it should be no surprise if our best scientific and philosophical understandings of the mind borrow heavily from it (and vice versa). It will be a very different world before scientists can do completely without thinking about what people want and believe.

Whether, however, philosophy and science can best profit from the raw, unwashed, folk concepts of belief and desire, or whether they should, instead, feel at liberty to modify and adjust these concepts, is another question. Indeed, folk intuitions about

belief may not all pull in the same direction or be entirely self-consistent. In such cases, at least, we should expect that philosophical and scientific investigations could profit from straightening and clarifying folk concepts to a certain extent. On the other hand, an account that strays too far from folk intuition risks losing insights from a long tradition of successful folk psychologists and may even lose justification for describing itself as an account of *belief*. I therefore aim to strike a balance between slavish adherence to intuition and sanctimonious disdain for it.

2. All-or-Nothing Belief and the Simple Question

The positive account to be given in the following chapter will be easier to accept if I first describe some of the intuitions that run against it, their utility, and where that utility ends. Doing so will, I hope, drain the power these intuitions might have to undermine my positive presentation.

The Simple Question

Most of us feel a certain temptation when presented with in-between cases of the sort that will be the focus of my account. The temptation is to insist on what I will call the *Simple Question* about belief (following Goldstein 1993). A person may be said to be asking the Simple Question about belief when two conditions obtain. First, she must be asking whether some thinking creature *S* believes some proposition *P*.[2] Second, she must accept only a simple yes-or-no answer to this question. One might think of an attorney cross-examining a hedging and evasive witness, saying, "Look, Mr. *X,* I am only asking you a *simple question,* Do you believe that *P* or not? Yes or no?" The idea behind insistence on the Simple Question is presumably that with enough tenacious probing, the evidence regarding *S*'s beliefs about *P,* evidence which may presently be tangled and indecisive, will eventually straighten itself out in favor of either *S*'s genuinely believing that *P* or *S*'s genuinely not believing that *P*.

[2] By 'proposition' here I simply mean 'candidate for belief' (cf. chapter three). Some term of art is needed here, since ordinary language provides no convenient term for such things. Nothing I say hinges on one or another resolution of the various metaphysical disputes about the nature of propositions.

Somewhere in the labyrinth of *S*'s mind, the reasoning goes, *P* has either set up residence, or it has not. Insistence on the Simple Question will not let us rest until we discover which is the case.

An inclination to insist on the Simple Question about belief has some very practical benefits. Suppose someone tells me that the notorious gambler Charlie Smart refuses to play poker without a package of salt in his pocket because, he says, the salt gives him good luck at the table, and suppose I have good reason to think, from other circumstances, that Charlie is a cool, unmystical probability theorist. My evidence regarding Charlie's beliefs on the topic of the effectiveness of lucky charms is now mixed. I could, at this point, simply assume that Charlie really is confused and inconsistent on the matter, or I could act on the hunch that there must be a resolution to this apparent tension and press the Simple Question: Does Charlie *really* believe that the salt will improve his chances? The inclination to take the latter route, to challenge evidence pointing in different directions regarding a person's beliefs, is a healthy one: Often there will be a perfectly good resolution of the tension. Charlie might not be as cool and unmystical as I thought. Perhaps even in the most serious vein he would avow the causal efficacy of lucky charms. Alternately, Charlie might *not* really believe in the efficacy of his charm. He is just sentimental, carrying the salt in memory of the last wish of his more mystical friend Idaho Bob who thought more highly of such methods.

Probing for a “yes” or a “no” in such a case may be helpful in eliciting an explanation of pieces of evidence pointing in different directions regarding an agent’s beliefs. Because of its utility, the inclination to insist on the Simple Question, at least when first presented with a tension of this sort, is nearly universal.

However, if I am told that Charlie, when pressed, repudiates with all sincerity belief in lucky charms but nevertheless becomes extremely uncomfortable and edgy, complaining of bad luck, if asked to gamble without his salt; if I am told that he is surprised when he loses carrying his salt and surprised when he wins without it, but regards his habit of carrying the salt as silly and superstitious -- if, in fact, a hundred different signs point in one direction regarding his belief and just as many point in the opposite direction, and there seems to be no hope of reconciling them -- it may be that Charlie is *not* accurately describable as either simply believing or simply not believing in the efficacy of his salt, and that insistence on the Simple Question will be counterproductive. One might just as sensibly insist on a simple yes-or-no answer to the question of whether Betty is courageous *simpliciter* when she is courageous in matters of love and money and cowardly in matters of health and work.

There is a limit, then, to the utility of insisting on the Simple Question. People are sometimes *not* accurately describable as simply either believing that *P* or not believing it. When it becomes clear that the case in hand is of this type, continued

insistence on the Simple Question becomes a hindrance rather than an aid to further research. The inclination to insist on the Simple Question, however, never entirely disappears. It is this continued inclination to insist on the Simple Question that I believe to be the most persistent source of dissatisfaction with the account of belief I will present. If I can succeed in motivating the reader to distrust this inclination, I will have gone far, I think, toward disposing the reader toward my account.

Several of my projects in chapters five through seven will, I hope, do something to motivate the reader to distrust any inclination she may have to insist too strenuously on the Simple Question. In the remainder of this section, I will describe and criticize the all-or-nothing view of belief implicit in refusal to abandon the Simple Question. In the following section, I will examine a pervasive metaphor in psychology and philosophy of mind that may be working to bolster our unwitting dependence on this all-or-nothing view of belief. In chapter six I will describe four areas in philosophy and developmental psychology in which too dogged an insistence on the Simple Question has led researchers astray. And throughout these chapters I will continue to provide detailed examples of the kind of in-between beliefs that do not fit into the categories allowed by the Simple Question.

The All-or-Nothing View of Belief

Only one view can justify unrelenting insistence on the Simple Question: the view that belief is inherently an all-or-nothing matter; for only if there can be no cases lying between full belief in *P* and complete lack of belief in *P* (or at least no cases we can be sure of) will insistence on the Simple Question always be appropriate. Few, I think, would want on reflection to endorse an all-or-nothing view of belief. We can see that the all-or-nothing view is not acceptable by examining three positions that follow from an all-or-nothing view of belief. I will sketch some widely accepted objections to two of these positions. I will also outline some concerns regarding the third position, to which I shall return briefly again at the end of my discussion of belief.

(1.) *Nonprobabilism*. The Bayesians are mistaken in saying there is a smooth gradation from indifference between *P* and not-*P* to certainty that *P* is the case, or from subjective probability .5 to subjective probability 1. If there are different degrees of certainty, they are only differences in one's attitude to propositions already completely and fully believed.

(2.) *Individuationism*. This view has two components: (1.) that beliefs are distinct and clearly individuatable, and (2.) that there is always a precise fact of the matter exactly which beliefs a subject has at any given time. If Mary is running upstairs to retrieve her purse from the bed, she may have some of the following beliefs: (a.) her purse is on the bed, (b.) her

purse is near where she slept last night, (c.) the object containing her lipstick is a few feet from the surface of the floor, etc. Individuationism commits one to the view that such beliefs are cleanly distinguishable and that there is a precise fact of the matter which of them Mary has and which she does not.

(3.) *Inaccessibilism*. A person who does not recognize in herself a belief that *P,* or who is cognitively incapable of acting on the basis of that belief in a certain range of circumstances, may still be said to believe that *P* as fully and completely as someone who does recognize that belief in herself and who can act on that belief in any circumstance. In the former case, the belief is genuinely present but simply "inaccessible" to the agent -- believed, perhaps, "implicitly" or "unconsciously".

Let us now consider these three corollaries of the all-or-nothing view. We have excellent reason to reject the first corollary, nonprobabilism about belief. We have, in other words, excellent reason to regard confidence about the truth of a proposition as the kind of thing that comes in degrees, spreading smoothly from indifference to absolute certainty. This view is so widely held that it is almost embarrassing to argue for it.[3] Jeffrey (1992) provides an elegant defense of probabilism, though his views are stronger than is needed here. Jeffrey claims that *all* our beliefs, even those sometimes taken as "foundational", are subject to the probabilistic calculus of Bayesianism; all

that is necessary for the rejection of nonprobabilism is that *some* of our beliefs are.

That belief comes in degrees seems quite plainly to be the everyday view, even if the everyday view does not quantify degree of belief. Someone can be absolutely certain, moderately sure, hesitant, doubtful, or cautiously accepting of *P*. The degree of confidence with which someone believes that *P* has a variety of effects recognized in folk psychology. The more confidently one holds a view, the more one is willing to stake on it, the less likely one is to revise it in light of counterevidence, the more forceful the conclusions one is willing to draw from it, the more assuredly one is willing to act on it, and the fewer hedges one will make against its falsity. And again, these generalizations from folk psychology seem smoothly extensible downward from the heights of confidence to the depths of uncertainty.

Bayesian decision theory, as elaborated by Jeffrey (1983), Ramsey (1990), Savage (1972), and others, builds upon these ordinary observations and quantifies them, generating a normative calculus for decision-making. Although decision theory is not free from difficulties, its range of successes would be hard to explain if it weren't right at least about the basic fact that beliefs are the kinds of things that come in degrees.

The second corollary to the all-or-nothing view of belief, individuationism, may seem more appealing than nonprobabilism. Suppose, for example, that one regarded beliefs as items in the

[3] Harman (1986) provides some reasons to think that nonprobabilistic full acceptance is our normal mode of dealing with propositions explicitly believed. But even Harman will

mind written in the “language of thought”, as Fodor does (Fodor 1975). If this were one’s view, individuationism might come naturally. If two purportedly identical beliefs correspond to the same sentence in the language of thought, they are the same belief; if they correspond to difference sentences, they are different beliefs; and it’s hard to see room for vagueness on the question of whether two sentences in the language of thought are the same or not (the first component of individuationism). There may be a little room on this view to deny there is always a precise fact of the matter exactly which of these sharply individuatable sentences are inscribed in a person’s mind (the second component of individuationism), but it strains against the model and images invoked. Fodor indeed may come closer than most to subscribing to an all-or-nothing or Simple-Question view of belief. He is also fond of the “belief box” metaphor I will discuss later in this chapter (Fodor 1987).

Individuationism, however, fares poorly on inspection. Holistic arguments are one natural avenue for criticism of this thesis. Suppose you and I both have a belief we describe by means of the sentence ‘Angela is fond of trees’. You, however, being unfamiliar with the proper meaning of the English word ‘tree’, take yourself to be expressing the belief that most of us would express with the sentence ‘Angela is fond of processed lumber’. You are agnostic about her attitude toward what we usually call trees. Clearly, we do not have the same belief on the subject. But change the case a little: You think of trees as

---

not go so far as to say that *all* our beliefs are nonprobabilistically accepted.

including both lumber and those living things (pines, redwoods, oaks, but not eucalyptus or orange) that are commonly turned into processed lumber. Now is it accurate to say that you share my belief that Angela is fond of trees? What if you don't think of any processed lumber as belonging to that class? What if you and I only disagree about the membership of saguaros in this class? What if you, like Davidson's dog (see chapter two) don't realize that trees require water and sunlight to grow? Presumably, if we share enough of our other tree-related beliefs, we will want to say that we share the belief in question, but when, exactly, is this line crossed? The difficulty of keeping facts about language and the expression of beliefs separate from facts about the beliefs themselves only adds to the confusion.

Generalizing from this example, it seems plausible to suppose that there is often a smooth spectrum of states between believing that *P* (Angela is fond of trees) while not believing that *Q* (Angela is fond of processed lumber) and believing that *Q* while not believing that *P*. It is not sensible to insist that a subject standing in the middle of this spectrum always be classifiable simply as believing that *P* or simply as believing that *Q*. Rather, in such situations, describing the subject's cognitive state as a belief that *P* or a belief that *Q* is somewhat a matter of approximation. The descriptions are more or less apt, not wholly accurate or wholly inaccurate. Individuationism requires the contrary, that one of the descriptions be exactly on

target and the other be a complete miss (even if it comes near to being a hit).

Another route to the rejection of individuationism about beliefs is suggested by the example I first gave in describing individuationism and which is borrowed from Dennett (1987, p. 111; Stich 1983 makes a similar case). This argument, like the previous one, depends on the implausibility of drawing a clear line across a smooth gradation. Where the previous argument depended on blurring the line between different propositions, the present argument concedes the existence of clearly individuatable beliefs and challenges the further claim that there is some precise fact of the matter which of these beliefs the subject genuinely has.

Consider Mary, then. Her date is waiting in the foyer. She is running upstairs to retrieve her purse. She believes that her purse is on the bed, which in fact it is. Mary would seem to have a number of related beliefs as well. She believes, for example, that her purse is in the bedroom. She believes her purse exists. Perhaps slightly more questionably, we can say that she believes her purse is near where she slept last night and that it is on some flat surface in her bedroom. Does she believe that her birthday gift from Allan is in the bedroom? Does she believe that her birthday gift from Allan is further from her date than she herself is? Does she believe that either her purse is in the bedroom or Fermat's last theorem is false? Does she believe that an object weighing 1.4 kilos is preventing light from reflecting off part of her bedcover? She herself will

answer yes to some of these questions, and no to others, depending on the context in which these questions are asked and the tone in which they are asked. Her intuitions on the matter waver. She would answer, if queried, that her purse is preventing light from reflecting off part of her bedcover, but she will deny having thought of it that way before. Surely we don't want to grant her belief in everything she would on (some sufficient amount of) reflection assent to -- but at the same time we don't want to assert that she believes only things that are presently passing through her consciousness. It is fantasy to think we can draw a strict line here between what she believes and what she does not. We should rather think of these descriptions as more or less appropriate for capturing Mary's cognitive state. Furthermore, the aptness of the descriptions will depend on the situation in which the description is provided. Individuationism, as I have characterized it, is false because there is no precise fact of the matter exactly which among a vast network of related propositions a person can accurately be said to believe. As in the lumber case, the appropriateness of describing a subject as believing a certain proposition seems to be a matter of *degree*.

Finally, let us consider inaccessibilism, the third corollary of the all-or-nothing view of belief. Inaccessibilism, as described above, is the view that a person who does not see herself as believing that *P,* or who is unable to act on *P* in all circumstances, might nonetheless be describable with the highest

degree of accuracy as believing that *P*. The belief that *P* might, in popular locution, really be "in there" somewhere, with the subject unable to access it for the time being. There is something rather intuitive about the inaccessibilist view. Perhaps I cannot now bring to mind, no matter how hard I try, the name of my sophomore year roommate in college. Still, I insist, I *know* his name. Or perhaps, though I deny it to myself, my pattern of behavior is generally racist. In such a case, we might say, I really do believe that one race is superior to another, but I cannot see that this is in fact my view.

I am happy to admit that it is more accurate to describe me, in these cases, as believing that my roommate's name was 'Louis' and that the caucasian race is superior, than it is to describe me as not believing these things; but it is a separate question whether it is *just as* accurate to ascribe me these beliefs as it is to ascribe them to someone who explicitly avows them. I think intuition pulls both directions on this matter. The impulses that drive us toward the Simple-Question, all-or-nothing view of belief incline us to say that, given that I do in fact have the belief in all these cases, there can be no "more or less" about it. The belief is really *in there,* and all the belief ascriptions are equally -- that is to say, 100% -- accurate. Nevertheless, people may feel at least some resistance toward saying that I do genuinely and completely believe, right now as I stand here stammering, that my sophomore year roommate's name was 'Louis'. And does it really seem completely accurate, in all

contexts, to say that I believe caucasians to be the superior race if I only believe it “deep down” and completely deny it on its face?

I ultimately want to reject the inaccessibilist view along with the nonprobabilist and individuationist views, but at this point in the presentation I will settle for a draw on the basis of conflicting intuitions. Inaccessibilism is incompatible with the account of belief I will present in the next chapter, and I aim to gather enough points in favor of my account of belief that it will seem reasonable to reject something as unstable as our inaccessibilist intuitions in favor of the picture I offer. I am, however, aware that this is a point on which my account might sometimes seem seriously to be at odds with intuition.

To review: The general thrust of this section is that it is quite natural, for good reasons, to insist on simple all-or-nothing answers in most inquiries about belief. Nevertheless, as I hope to have made plain, the all-or-nothing view of belief is untenable for a variety of reasons. I shall now move on to describe a metaphor commonly used in talking about the mind that may also be partly responsible for leading us unreflectively into thinking of belief as an all-or-nothing matter.

3. The Container Metaphor

A metaphor is a powerful force, and the persistent use of any particular metaphor inevitably draws its users' thoughts in a certain direction. Lakoff and Johnson examine, for example, the regular metaphorical treatment of arguments as battles: Arguments are won or lost; positions are attacked and defended, shot down or salvaged; criticisms are launched and found to be on or off target; and so forth (1980 p. 4). They argue that this way of talking about argument is apt to influence one's thinking about and approach to argumentation, making one, perhaps, more combative in one's argumentative style and less likely to notice the co-operative aspects of argumentation.

Much of our talk about the mind is likewise metaphorical, both in everyday discourse and in technical philosophy and psychology. As with our metaphors for argumentation, the metaphors we use to talk about the mind doubtless incline us to think of the mind in one way rather than another. It would therefore seem to be of extraordinary importance in a discussion of how to think of the mind to examine the metaphors we employ in talking about it. Unfortunately, this is rarely done.

In this section I will examine one persistent metaphor in philosophy of mind and indicate how its use might incline one toward the all-or-nothing view of belief and other disputable doctrines about the mind. I do not mean to claim that everyone who employs this metaphor holds the views suggested by the metaphorical usage. Metaphor is not destiny. But I do think

that these positions have a certain attractiveness they might not otherwise have in virtue of being suggested by the popular metaphor, and I hope for two effects from displaying this metaphor as a source of their attractiveness. First, I hope that revealing the metaphor as a source of attraction helps to bring more acutely into question the reasons people might have for being inclined toward these positions. Second, I hope that revealing some of the directions in which this metaphor leads our thinking will incline us to use the metaphor less frequently and with greater awareness.

Lakoff and Johnson (1980) have some useful discussions of metaphors used in talking about the mind. They discuss, for example, the metaphor of the mind as a machine (grinding out solutions to problems, feeling rusty, running out of steam), of the mind as a brittle object (I am going to pieces, her ego is fragile, he is easily crushed), and of ideas as food (half-baked), plants (coming to fruition), commodities (to be packaged), and fashions (out-of-date) (1980 p. 27-28, 46-48). They also very briefly mention, although they do not provide any examples of, the metaphor that will be the focus of my attention: the mind as a container or storage space (p. 148).

That the metaphor of the mind as a container is commonly used in everyday discourse can be made clear by a few examples:

> He *filled my head* with new ideas.
> Keep that thought *in mind*.
> Don't *clutter up your mind* with that *rubbish*.
> He *crammed* for the exam.
> Memory *retrieval* can take effort.
> *Empty your mind* of thoughts.

That person sure is *airheaded*.

The container metaphor in cognitive psychology is often quite explicit in discussions of memory *storage* and *retrieval*. In philosophy of mind, the prevalence of the container metaphor is most apparent in the popularity of the word 'content'. Interestingly, talk about "mental contents" takes place on two levels at once: Minds are said to have contents, of which beliefs and desires are of course the most popular examples (some, such as Fodor (1987), even talk about "belief boxes"); at the same time, beliefs and desires are themselves said to have "propositional contents". It is primarily on the first of these container relations that I will focus my attention, though I do not doubt that discussions of propositional content could also profit from a more scrupulous look at the metaphors involved.[4]

This metaphorical treatment of the mind as containing beliefs is appropriate if the relationship between minds and beliefs is similar in important ways to the relationship between prototypical containers and their contents. Even if the mind is viewed *literally* as a container for beliefs, presumably the extension of the class 'container' to cover minds is warranted only if there are such similarities. The same holds for the view that containers provide a good *model* of the mind. Even, then, if one were to argue that philosophical or psychological reference to containers in discussing the mind is not metaphorical, proper use of container talk depends on the existence of similarities

[4] Reddy (1979) and Lakoff and Johnson (1980) have interesting discussions on the related metaphor of linguistic expressions as having propositional content.

between paradigm instances of containment and relations into which the mind can enter.

This is certainly not to say that proper talk of the mind as containing beliefs requires that the relationship between minds and beliefs must be in every respect like prototypical instances of containment. If I say that Richard is a tortoise when it comes to paying his bills, I do not mean to be suggesting that Richard's skin is scaly or that he carries a hard shell on his back, and no one with a standard, American cultural background would regard me as suggesting this (though one could imagine strange enough contexts in which this could be the meaning). A somewhat more elaborate example is the planetary model of the atom, invoked metaphorically in talk about electrons *orbiting* the nucleus. (I am intentionally blurring here the difference between a model and a metaphor; I actually believe that the differences are less than they are sometimes supposed to be; for a good discussion see Black (1962).) Although this model is still frequently used in explaining the structure of the atom, especially in teaching, it has several infelicities which, if not made clear, can hamper understanding. The atom is like a planetary system in that it has a large mass at its center, several smaller masses that maintain themselves at a distance from it, a lot of empty space between the masses, and so forth. On the other hand, planets have definite positions in space, while electrons are thought to be "spread out" over an area, planets make regular elliptical orbits, while measurements of

electron position yield less regular results, and so forth. Once a model or metaphor is in place, especially if it is repeated frequently, the mind will naturally attempt to extend it in plausible directions, and students employing the planetary model of the atom must be specifically warned against these inferences. Black describes both the power and danger inherent in this tendency to draw inferences from models of this kind.

It is my belief that the container view of the mind has many more infelicities than advantages. We can discover problems even at the most basic ontological level. Objects are not usually thought of as containing their states, but beliefs and desires are generally regarded as states of minds. So the view that minds contain beliefs and desires seems to rest on a category mistake, like regarding *being 17° Centigrade* as something a bucket contains because it is in that state.

Although that ontological matter is worrying, it is not my primary concern. After all, if the container metaphor is apt in other ways, one can always warn against particular inferences. I will turn my attention to more the more specific features of prototypical instances of containment. Enough of these features are inappropriate to the mind-belief relation that the container metaphor for the mind has substantial potential to mislead. Of particular interest for my overarching project are those features of containers that suggest the all-or-nothing view of belief, but I will not confine my list of features to those suggesting that view.

For concreteness, I will take upright buckets as prototypical containers. I will also regard (discrete, undivided) balls as the contents (see footnote five for a discussion of liquid contents). If it is useful to think of the mind as "containing" beliefs, then the mind should be, at least in some important respects, like the bucket, and individual beliefs should be like the balls. I shall now describe some of the relevant features of the bucket-and-ball system.

(1.) A bucket contains a ball just in case the ball is physically inside the bucket. In other words, the volume of the ball must be a part of the volume enclosed by the bucket. It does not matter how things stand outside of the bucket.

(2.) In the normal (upright, gravitational) case, it takes a certain amount of effort to get a ball into a bucket and a certain amount of effort to get it back out again.

(3.) Balls take up space. A finite bucket can only contain a limited number of non-infinitesmal balls. It takes a certain amount of the bucket's spatial resources to contain each ball it contains.

(4.) Balls are typically clearly individuated, countable entities. We can, of course, imagine cases in which this is not so: Rubber balls may be melted together, balls may be cut into pieces, etc.; but these are not the kinds of things that typically come to mind when we imagine container relations between buckets and balls.

(5.) A ball is generally either fully inside a bucket or fully outside it. In marginal cases, a ball may be suspended

near the lip of a bucket, or it may be unclear for reasons of topology whether its volume is part of the volume enclosed by the bucket; also, of course, as balls enter and leave buckets there will typically be a brief period during which they may be said to be neither wholly inside nor wholly outside the bucket. Despite these marginal cases, however, it is rarely a vague matter whether a bucket contains a ball or not.

(6.) If the balls are small enough and appropriately shaped (and not, for example, highly magnetized), there is typically no reason why any two balls can't go in the same bucket or why a ball can't be removed from one bucket and put into another without changing any of the other contents.

(7.) A bucket can contain only one ball, or no balls.

Just as the argument-as-battle metaphor naturally inclines one toward a certain view of argumentation -- a view one might, on reflection, want to reject -- so, I would suggest, the mind-as-container metaphor, in virtue of the features described, naturally draws one toward a certain view of belief. The view of belief toward which we are drawn by the container metaphor has a number of undesirable, or at least controversial, features.

If the mind-belief relation has the features described in (1.), wherein the containment of a thing depends only on that thing's being inside the container, then beliefs must be things *internal* to the mind, *contra* the externalist view, to be discussed in the next chapter, of beliefs as partly dependent on

social or historical relations between the subject and the external world.

The features described in (2.), regarding the effort involved in adding and removing objects from containers, do not sit comfortably with our knowledge of how hard it can be to remember things and how easy it can be to forget them.

If the mind-belief relation has the features described in (3.), wherein buckets are characterized as containing only a limited number of balls, then we can only have a limited number of beliefs. Many have argued, however, that the number of beliefs any person may have is indefinitely large, since, it seems, I believe that the number of planets is less than 10, I believe that the number of planets is less than 11, and so on upward (see, for example, Harman 1986; Dennett 1978).

If (4.), the claim that balls are clearly individuatable, captures a feature of the mind-belief relation, then beliefs must also be clearly individuatable; and combining (4.) with (5.), the under which balls are either fully inside or fully outside a container, suggests that there must be a precise fact of the matter exactly which of these beliefs a subject has at any given time. These two combined, then, suggest “individuationism” as described in the previous section.

Furthermore, (5.) taken alone suggests also suggests the doctrine of “nonprobablism” as described in that section.

The sixth and seventh features of containers, relating to the independence of the presence of one ball in a bucket from the presence of others balls, are incompatible with a holistic view

of belief on which the possession of any single belief is dependent upon and changeable with the possession of many other beliefs.[5]

That so many of these features of the container relation seem, at least to some people, not to apply to the relation between the mind and beliefs is testimony to the fact that the use of a metaphor does not commit its user to regarding the object described metaphorically as having all of the features the metaphor suggests. But let us not slip into thinking that the metaphor is completely innocuous. Repeated application of the container metaphor is bound to pressure us subtly into certain habits of thinking, though we may successfully resist it in our more reflective moments. We should aim to be especially careful in examining the justification of positions suggested by such metaphorical uses. People with a particular interest in rejecting the patterns of thinking that come with a metaphor may wish to avoid at least that metaphor's livelier uses.

We ought, then, to be wary of letting talk about mental content lead us unreflectively into treating any of the features

[5] The metaphor can be extended or the model adjusted with an eye to avoiding at least (5.) above. The bucket is again the mind or the believing faculty of the mind. The beliefs, instead of being balls, are different liquids. The amount of liquid *P* contained in the bucket corresponds to the subject's degree of belief that *P* is the case. This model does avoid the nonprobabilism suggested by the earlier model, but (1.) - (3.) and (6.) - (7.) still clearly apply. One might try to get around (4.) by noting that different mixtures of liquids are not clearly individuatable, but the maneuver fails: A mixture of *A* and *B,* once in the bucket, is indistinguishable from *A* and *B* added separately, but these two cases must be kept distinct if the model of overlapping, not clearly individuatable mixtures is to have any value. The chemically pure liquid is thus the natural unit of analysis, and chemically pure liquids are neatly distinct from each other.

Other changes may of course be introduced. To avoid some of the more obvious difficulties with (2.), one might imagine the bucket having a spout through which old balls are pushed as new balls are added. Or, *contra* (6.), balls may be imbued with properties that make it difficult for a bucket to contain certain of them simultaneously, and so forth. There is sufficient material here for hours of fun. The point remains, however, that until such changes are actually introduced into our way of talking about

following from this metaphor as features of the mind, unless we can provide independent reasons for accepting those features. The first images that come to mind when the container metaphor is invoked are just as apt to mislead than to clarify.

beliefs, the more basic metaphor is the one that will have the greatest impact on our way of thinking.

4. Conclusion

In chapter seventh I will offer several in-depth examples of arguments in philosophy and developmental psychology which seem to suffer from an unreflective treatment of belief as an all-or-nothing matter. To what, exactly, we should attribute the tendency to overlook the possibility of in-between states of believing is not a matter I can hope to have settled. I have in this chapter offered what I regard as two plausible explanations: that the natural advantages of insisting on the Simple Question may lead us to take this insistence too far; and that steady repetition of the container metaphor may incline us, at least in our less guarded moments, toward thinking of belief as an all-or-nothing matter. In the next chapter I will describe a view of belief that recognizes the importance of in-between states of believing and invokes a metaphor much friendlier to matters of degree than is the container metaphor.

Chapter Six

A Phenomenal, Dispositional Account of Belief

In this chapter, I offer what I call a *phenomenal, dispositional* account of belief. I call it a *dispositional* account because it treats believing as nothing more or less than being disposed to do and to experience certain kinds of things. I call it a *phenomenal* account because, unlike dispositional accounts as typically conceived, it gives a central role to first-person, subjective experience, or "phenomenology."

Dispositional accounts are usually thought to be motivated by a desire to justify talk about mental states by reducing it to talk about something that behavioristically-minded philosophers find less objectionable, *viz.* dispositions to behave. I want to make it clear from the start that this reductionist motivation plays no role my project. My aim in presenting this account is, as I hope became clear in the previous chapter, to describe a way of thinking about belief that is both faithful to the facts and useful for the purposes of philosophy and psychology — an account, especially, that can provide us with a framework for understanding subjects not accurately describable as either simply believing that *P* or simply not believing that *P,* subjects in what I have called *in-between* states of believing. It is not necessary for this purpose — in fact, it is positively

detrimental — to insist on reducing mental state talk to talk about anything else.

I will begin with a statement of the account. I will then discuss in-between states of believing in some detail. I will conclude with a discussion of the relations between the present account of belief and several other positions one might take regarding belief.

1. The Account

It will be helpful to begin by disarming several preconceptions the reader may have about accounts of belief that focus on the dispositions of the believer. I have already alluded in the introduction to two of these preconceptions.

First, as I suggested in the introduction, a dispositional account of belief need not aim at being *reductive*. It need not, in other words, aim to show how all talk about beliefs (in particular) or mental states (in general) can be transformed or "reduced" into talk about other, less objectionable things. It is rare in science to manage reductions of this sort, in which a whole range of discourse is shown to be replaceable by some other different kind of discourse. Fortunately, insight into scientific subjects does not seem to require such reductions. In describing the dispositions relevant to a belief, I will feel no compunction about appealing to dispositions that themselves involve beliefs. So, for example, relevant to Maurice's belief that smoking is dangerous is his disposition to recommend against it, if he believes that the recommendation will do any good.

A second preconception about dispositional accounts of belief is that they can only appeal to *behavioral* dispositions. Once a dispositional account of belief is unshackled from reductivist demands, however, the range of allowable dispositions broadens substantially. Dispositions to acquire new beliefs and desires, for example, would be perfectly acceptable. Especially important, in my view, are what I will call *phenomenal*

dispositions — dispositions, that is, to undergo certain kinds of subjective, phenomenal experiences, like a conscientious student's disposition to feel surprise and disappointment were she to get a B- on a paper. In calling my account a *phenomenal* dispositional account, I mean to be emphasizing the role these phenomenal dispositions play in belief.

A third preconception about dispositional accounts of belief has to do with what it is to have a disposition. Ryle, who launched contemporary interest in dispositionalism, made a point of arguing that dispositions are bare 'inference tickets,' licensing us to make hypothetical claims of the sort, "If P occurs, then Q will," but in no way warranting inferences about the existence of any non-dispositional states or facts underlying the dispositions in virtue of which the dispositional claims are true (Ryle 1949). Ryle's account of dispositions has since been the subject of much critical scrutiny (for a review, see Prior 1985), and there is no need to attach his particular view to dispositional accounts of belief in general. My dispositional account of belief is in fact quite compatible with a robust, anti-Rylean view of the physical and causal underpinnings of dispositional properties.

My account of belief employs the concept of a *dispositional stereotype* for a belief. The notion of stereotype to which I am appealing here is somewhat like that described in Putnam

(1975a).[1] A stereotype is a cluster of properties conventionally associated with a thing, class of things, or property. To use Putnam's example, stereotypical properties of tigers include their being striped and their being four-legged. Some things worth being classified as tigers — tigers dipped in ink, three-legged tigers — may not have all the stereotypical features of tigers; although such creatures may be tigers, they are not stereotypical ones. Indeed, we might discover that some of the stereotypical features of tigers are had by no tigers at all (for example, if it were part of the stereotype of tigers that they lived in African jungles). Stereotypes may in fact be broadly inaccurate, although this is not normally the case. Putnam points out that the stereotype for gold involves its being yellow, although chemically pure gold is more nearly white.

Understanding dispositional stereotypes also involves understanding *dispositions*. Prior (1985) again provides a useful review of contemporary positions. Without getting overly involved in the tangle of issues arising in the philosophical debate on the nature of dispositions, I would characterize a disposition by means of a conditional statement of this form: If condition *C* holds, then object *O* will (or is likely to) enter (or remain in) state *S*. *O*'s entering state *S* we may call the *manifestation* of the disposition, and condition *C* we may call the *trigger* or *condition of manifestation* of the disposition. Exactly what the connection is between *O*'s having the

[1] The present concept of 'stereotype' does differ from Putnam's in associating stereotypes with things rather than with words, and in seeing it as a cluster of

dispositional property to enter state *S* in condition *C* and the truth of the conditional statement associated with that disposition is a matter of some debate, but as a rule of thumb, we may suppose that *O* has the disposition in question just in case the corresponding conditional statement is true. Thus, for example, salt has the dispositional property of being soluble in water because it is apt to dissolve (the manifestation) when placed in water (the trigger). Mirrors are disposed to reflect light because when light shines on them (the trigger), it reflects back (the manifestation). Carlos is disposed today to get angry when his car doesn't start because if his car doesn't start today, he is likely to get angry.

A *dispositional stereotype,* then, is a stereotype whose elements are dispositional properties. Consider, for example, the stereotype for *being a reliable person*. This stereotype will include the disposition to show up to meetings on time, the dispositions to follow through on commitments, to be prudent and careful in making important decisions, and so forth. Personality traits, such as being hot-tempered, courageous, tenacious, and so forth, are all characterizable by means of such dispositional stereotypes. To have these personality traits is really nothing more than to match these stereotypes. My core claim is that belief can be characterized in much the same way.

Thus, consider a favorite belief of philosophers: the belief that there is a beer in the fridge. A sample of the dispositions associated with this belief includes: the disposition to utter,

---

properties rather than as a set of ideas.

in appropriate circumstances, sentences like ‘There’s a beer in my fridge’; the disposition to go to the fridge if one wants a beer; a readiness to offer beer to a thirsty guest; the disposition to think to oneself, in appropriate contexts, ‘There’s a beer in my fridge’; an aptness to feel surprise should one go to the fridge and find no beer; the disposition to draw conclusions logically entailed by the proposition that there is beer in the fridge (e.g. that there is something in the fridge, that there is beer in the house); and so forth.

It is important to notice that no *one* of these dispositions is either necessary or sufficient for the possession of belief. Intuitively, it may seem that the disposition to feel assent to an internal utterance of *P* comes close to being a sufficient condition for believing that *P;* nevertheless, we must allow that people sometimes feel assent to utterances that it is not wholly accurate to describe them as believing, e.g., when they don’t really understand what the utterance means or when they are “self-deceived.” (I will discuss the case of self-deception in chapter seven.)

The list of dispositions that informed common sense is capable of associating with any given belief may be indefinitely long. I would not want my talk about “stereotypes” to suggest that we must already have associated with each belief each of these dispositions. Rather, think of the dispositional stereotype for the belief that *P* as consisting of the cluster of dispositions that we are apt to associate with the belief that *P*.

These stereotypes will be composed primarily of behavioral and phenomenal dispositions, although other sorts of dispositions, such as dispositions to acquire new beliefs and desires, will play a role as well. The dispositions belonging to stereotypes for belief will include all the behavioral and other dispositions typically referred to by those advocating standard "functionalist" accounts of belief (Putnam 1966; Lewis 1972, 1980; Fodor 1968), as well as many phenomenal dispositions that play at most a derivative role in standard, functionalist accounts — such as dispositions to feel surprised or disappointed and to make internalized utterances.

The reason I say that the stereotype consists of a *cluster* of dispositions is to bring out two ideas: first, that some dispositions are more central to the stereotype than others, and second that there may be vagueness and conflict regarding exactly which among the more peripheral dispositions should belong to the stereotype. Stereotypes are not thereby rendered useless: Rosch (1977) and Wittgenstein (1958) have argued that many of our most useful concepts depend on clustering properties together in this way.

A person who possesses all the dispositions in the stereotype for believing "There is a beer in my fridge" can always, on my view, accurately be described as having the belief that there is a beer in his fridge. A person who possesses none of the relevant dispositions can never accurately be so described. And, of course, bridging the gap between these two extremes is a wide

range of cases in which the subject has some but not all the dispositions in the stereotype. Roughly speaking, the greater the proportion of stereotypical dispositions a person possesses, and the more central those dispositions are to the stereotype, the more appropriate it is to describe him as possessing the belief in question. An additional element of vagueness is introduced if one accepts that having a disposition is itself not a simple yes-or-no matter.

To believe that *P,* on the view I am proposing, is nothing more than to match to an appropriate degree and in appropriate respects the dispositional stereotype for believing that *P.* The *belief* that *P,* in any organism, is whatever state of that organism that causes it to respond in ways that match the dispositional stereotype for believing that *P*.[2] What respects and degrees of match are to count as "appropriate" will vary contextually and cannot be specified by any simple rule, and so must be left as a matter of judgment. I hope the numerous examples in this chapter and the next will help reveal what course such judgments tend to take. The view offered here does not imply, nor is it intended to suggest, that beliefs are

[2] An organism may then be said to "have a belief" just in case that organism is in a state that causes it to respond in ways that match the relevant dispositional stereotype. It is thus logically possible, on the definitions I have given, to believe that *P* but not to have the belief that *P* — if the organism matches the stereotype for believing that *P* but is not caused by any of its states to respond in the stereotypical ways. In a richly causal universe such as our own, however, believing that *P* and having the belief that *P* will always go hand in hand. If one is nonetheless concerned to close the logical gap between the characterizations I have given here, one might wish to alter the first sentence of the paragraph in the following way: To believe that *P* is to be in a state that causes one to respond in ways that match, to an appropriate degree and in appropriate respects, the dispositional stereotype for believing that *P*. I have no serious objections to such a definition of belief, although I think the definition in the text is simpler and for all practical purposes amounts to the same thing.

metaphysically secondary or that talk about them is somehow eliminable.

If a metaphor for talking about belief is necessary, I would prefer the metaphor of matching profiles to the container metaphor: Rather than thinking of *P* as the *content* of the belief that *P,* I would prefer to think of *P* as the *profile* of that belief. This allows, much better than the container metaphor does, in-between cases of the type that will shortly be occupying our attention. One's dispositions may have something of a *P*-ish profile, something of a *Q*-ish profile, or something in between; one's dispositional profile may match up quite precisely with *P* or it may be less exact a match. For a discussion of the infelicities of the container metaphor of belief, the reader is referred to the previous chapter.

Ceteris Paribus Clauses and Excusing Conditions

A substantial complication arises from the fact that common sense regards all these dispositions as holding only *ceteris paribus* or "all else being equal." Joe might believe there is beer in his fridge, but if he is particularly stingy with his beer, he may not have some of the dispositions described above — he may not, for example, be ready to offer a guest a beer or even to admit that there is beer in his fridge at all — but we wouldn't want to say that lack of these dispositions makes it any less accurate to describe him as having that belief. Behavioral dispositions seem particularly defeasible in this way, phenomenal

dispositions a bit less so: If we were to imagine that Joe was not disposed to feel surprise upon opening the fridge and noticing a lack of beer, this would generally seem to reduce at least to some degree the aptness of describing Joe as believing there is beer in his fridge.

In any case, the dispositions in the stereotype of a belief are best seen as defeasible, loaded with tacit "if" clauses, e.g., Joe is disposed to assent to utterances meaning that there is a beer in the fridge *if* he hears the utterance, *if* he has decided not to lie about the matter, *if* he understands the language in which the utterances take place, *if* he has the physical capacity to indicate assent, and so forth.

Note that in being *ceteris paribus* defeasible these dispositional claims are not different from many scientific and ordinary generalizations. Human beings are born with two legs *if* they have developed normally in the womb, *if* they don't have an unusual genetic make-up, *if* the doctor does not saw off a leg before removing the child, etc. Rivers erode their outside bank at a bend *if* the river is not frozen, *if* the bank is made of an erodable material, *if* there isn't a powerful fan in place preventing the water from touching the outside bank, etc. The *ceteris paribus* nature of such generalizations does not in these cases, nor I think in the dispositional case, hinder their productive use.

I leave it as an open metaphysical question whether the dispositions in question must always be manifested if all their

conditions of manifestation are met. If so, then dispositions must often have an indefinitely large number of tacit conditions: Condition *C* of the disposition's conditional characterization must, if completely fleshed out, be an indefinitely long conjunction. (I am presuming we do not want to cut the matter short by adding something like "and nothing prevents it" to the conditions of manifestation.) On the other hand, one may wish to include only a few conditions in the trigger for any given disposition, if one is not averse to the idea that dispositions do not always manifest themselves when their conditions are met (see Martin and Heil 1996). Attempting to resolve such questions would lead us away from our main project, since nothing in my account depends on such details.

A person may then be *excused* from a dispositional manifestation — i.e. not seen as deviating from the dispositional stereotype — if one of the tacit conditions of manifestation is not met or if the disposition is simply not manifested for some reason consistent with possession of the disposition, perhaps because it is blocked by another disposition. Certain types of conditions are regularly regarded as excusers, such as physical incapacity or the presence of a desire or situation that makes a particular manifestation prudentially inadvisable. If Joe's mouth is sealed shut, it does not count against his believing that there is beer in the fridge that he is unable to tell us so. Other conditions may be somewhat less excusing and are apt to propel us again into vagueness: ignorance about related topics

(e.g., Joe believes that Budweiser is not a type of beer), distraction by other cognitive demands, or apparent failure to reason correctly. If Joe knows there is only Budweiser in his fridge, but Joe does not think Budweiser is a type of beer, does Joe believe there is beer in his fridge? Those fond of the *de re/de dicto* distinction might remark that Joe seems to believe *(de re)* of a certain type of beer that it is in his fridge, but not to believe *(de dicto)* that there is beer in his fridge. This is only one way (and a questionable one: see Stich 1983; Dennett 1987) of trying to get a handle on intuitions that pull us in different directions in such cases.

One wants to find a single, unifying principle that can guide us in distinguishing cases of genuine deviation from excused non-manifestations. This is essentially a demand for a principle unifying all the *ceteris paribus* excusers from dispositional manifestation. I think the prospects for finding such a principle are slender, but a brief look at the question is nonetheless instructive.

Let us begin with examples. Certainly when there is a sense that the disposition in question would have manifested itself but for the presence of some hindrance external to the agent's mind, we are ready to grant excuses. If Joe doesn't offer beer to a guest only because someone with a gun to his head is telling him not to, we are hardly inclined to count his not offering beer as a mark against the accuracy of describing him as believing there is beer in the fridge. A general shutdown of the mind also seems

to be excusing: We don't blame Joe for not offering the beer if he has blacked out. On the other hand, if Joe denies having beer in his fridge when a guest requests some, and we cannot tag his denial on any external cause, nor on an intention to lie, nor on a misunderstanding of the question, there may be no explanation left other than to say he doesn't realize that there is beer in the fridge; if then, five minutes later, he turns around and offers his guest a beer, though nothing in the situation seems to have changed, we may be greatly puzzled. We look for some way to explain this "inconsistent" behavior: Perhaps he suddenly remembered there was beer in his fridge after all? What, then, are we to say about his belief five minutes ago — that he really did believe there was beer in his fridge, but only "in some corner of his mind"? Does it matter whether he would have recalled it then, had he only stopped to think more carefully about it? Even, however, if *some* of Joe's dispositions five minutes ago accorded with the stereotype, Joe's deviation from the stereotype at that time may have been symptomatic, in a way the deviations introduced at the beginning of this paragraph were not, of a systemwide likelihood of deviation from many aspects of the dispositional stereotype.

This last point may seem to hold some promise for the construction of a general principle differentiating excused non-manifestations from genuine deviations. In cases of linguistic misunderstanding, or of deliberate concealment, or of yielding to external pressures, failure to manifest the stereotypical

disposition does not seem to be symptomatic of a systemwide, behavioral and phenomenal nonconformity to the stereotype. Joe might well be thinking silently to himself, in any of these cases, “There is a beer in my fridge.” We have no reason, in such cases, to expect a *general* non-adherence to the stereotype; there seems to be a natural containment of the deviation to a particular range of circumstances: If the gunman were to walk away, if his guest were to start speaking English, if Joe did not feel his precious beer threatened by the presence of a thirsty guest, we would again see a general conformity to the stereotype. One could even bring cases of general mental or physical shutdown under this umbrella, if one were to think of these conditions as particular, narrow circumstances. Perhaps, then, some idea of containment of the deviation could be drafted to serve as a general principle for identifying excusing conditions.

The question then arises, however, whether in putting forward such a principle we have added anything of substance to the account. Scientific and everyday generalizations are shown false by deviations that undermine our reasons for thinking the generalization is widely, approximately, or at least in “ideal” circumstances, right; we introduce *ceteris paribus* excusers in just those cases where we feel that a deviation from the generalization does not affect its overall validity. Introducing a rule, then, that says *ceteris paribus* excusers are to be admitted exactly when a deviation does not threaten the basic accuracy of the generalization is simply to state what is

implicit in the admission of *ceteris paribus* defeasibility from the beginning.

Clarifying this point helps us to see the two factors that come together in assessing deviations as potentially excused. The first factor is an empirical assessment of the likelihood of the generalization's broadly falling apart given that the deviation has taken place. The second is a practical understanding of the role of the generalization in one's cognitive structuring of the world. Where can one afford a certain amount of looseness in the generalization because the cases are marginal or covered by other generalizations, and where will one want to insist on a stricter adherence to the rule? No set of explicit rules seems to be able to guide us as well in making these assessments as does a well-practiced intuitive grasp of the generalizations in question. This lack of explicitly specifiable rules for separating excused from unexcused deviations from a generalization infuses even the most robust scientific theories (for examples in physics, see Cartwright 1983). Philosophers of science have learned to resist the temptation of attempting to spell out in full detail the *ceteris paribus* conditions for any substantive, specific scientific generalizations.

A failure to manifest a disposition, then, can either be excused or unexcused. When the failure is excused, the deviation detracts not at all from the accuracy of describing the person in question as having the belief. When the lack of manifestation is

not fully excused, the question of whether it will count as an *important* deviation — one that makes us hesitate to ascribe the belief or makes the belief ascription less apt than it could be — will generally depend on the context in which the belief ascription takes place. Suppose, for example, that a child studying for a test reads, "The Pilgrims landed at Plymouth Rock in 1620," and remembers this. She is bit confused about what pilgrims are, though: She is unsure whether they were religious refugees or warriors or maybe even some kind of bird. Now, does she believe that the Pilgrims landed at Plymouth Rock in 1620? In some contexts — e.g., if we are talking about her likely performance on a history dates quiz — we might be inclined to describe her as believing this; in other contexts we would not. Note that I am not saying that the mental state of the child varies with context. Rather, given that the child deviates from the stereotype in some respects but not in others, how best to *describe* her mental state will depend on the practical demands of the moment.

This context-dependence is an important feature of the proposed account. Different dispositional properties will, in different contexts, be more or less crucial to decisions about whether to ascribe a particular belief or not, and in mixed cases failure to attend to the context of ascription can result in differing assessments of the appropriateness of a belief ascription. Such inattention to context may be partly responsible for much of the wavering and disagreement about how

to describe the kinds of in-between cases that are puzzling to those who approach these cases looking for all-or-nothing answers. Further examples of context-dependence in belief ascription will be developed as the discussion of belief continues.

The Importance of Phenomenology for a Dispositional Account

Dispositional accounts of mental states are not, of course, new. Gilbert Ryle's *The Concept of Mind* (1949) began a trend toward regarding much of mental life as fundamentally dispositional or at least as dispositionally specifiable (Armstrong 1968) — or, not so differently, as "functionally specifiable" (Lewis 1972, 1980; Putnam 1966; Fodor 1968). (A dispositionally specifiable state is a state of an object, e.g., a brain, apt to bring about specified effects under specified conditions; a functionally specifiable state is a state of an object apt to bring about specified effects under specified conditions and to be produced by specified causes.) Others have argued for dispositional accounts specifically of belief, or specifically of unconscious, non-"occurrent" belief, independent of any broader dispositionalist or functionalist program (e.g., Searle 1992; Marcus 1990). None of these accounts (except perhaps Searle's, which is in any case limited to unconscious beliefs), however, appeal to *phenomenal* dispositions in their characterizations of belief.

The inclusion of phenomenal dispositions in my account ensures that the standard anti-behaviorist objections to Ryle's dispositional account of belief are inapplicable. The most compelling of these objections belong to a single genus, exploiting the loose connection between mental states and behavior (e.g. Chisholm 1957; Putnam 1963; Strawson 1994). Putnam, for example, imagines a society of "super-spartans" who feel pain but do not exhibit the range of behaviors typically associated with pain (except avoidance, which is not specific to pain). Similarly, Strawson imagines a species of "weather watchers" who have beliefs and desires about the weather but are not constitutionally capable of acting in any way on the basis of those beliefs and desires. Chisholm emphasizes that we should not describe someone as disposed to act in a certain way, given a particular belief, unless we grant that that person has *other* particular beliefs and desires. For example, though Jones may have the belief that his aunt will be arriving at the railroad terminal in twenty-five minutes, it is only true to say he is disposed to go there to pick her up if he wants to pick her up and if his beliefs about how to get to the railroad terminal are not too deeply confused. Full conditions for the possession of any particular belief or desire can never be given in terms of behavioral dispositions alone; appeal to some other aspect of the subject's mental life will always be necessary.

The appeal to phenomenal dispositions gives the dispositionalist about belief a clear and natural way around

these objections. Putnam's super-spartans and Strawson's weather watchers, though they lack the manifest behaviors associated with believing, still have the *phenomenal life* attending belief — if they did not, there really would be no reason to regard them as believing. Furthermore, they have clear, typical excusers from behavioral manifestation: contrary desires in the case of the super-spartans and incapacity in the case of the weather watchers. We can also grant Chisholm his point: There is no way to analyze away mental life in favor of behavioral dispositions or to replace all talk of belief with some other kind of talk. These are behaviorist aims not naturally suited to a non-behavioralist dispositionalism. Since it is no part of phenomenal dispositionalism to bring about these ends, it is no objection to phenomenal dispositionalism that it is impossible to do so. Nevertheless, it is interesting to observe that at least some phenomenal dispositions have quite a tight connection between trigger and manifestation. A person who believes that *P* will normally feel assent to an internal utterance or verbal image of a sentence expressing *P* in her own language *regardless* of what else is true of her; similarly for her feeling surprise at discovering that *P* is false. If she is not disposed to feel assent toward the thought that *P* or feel surprise at finding *P* false, we rarely allow excusers: These are central cases of deviation from the stereotype.[3] (We may nonetheless want to ascribe the belief if the subject matches the stereotype in

[3] Assuming that a person has privileged access to her own phenomenology, we may have here the beginnings of an explanation of the high accuracy of first-person belief

enough other respects, which is part of why feeling assent to an utterance of *P* is not equivalent to believing that *P*.)

A Thought on Ryle

I would like to conclude this section with some remarks about Ryle, the intellectual forefather of dispositionalism about mental states. Although he is typically viewed as a behaviorist for whom appeal to phenomenal dispositions would be strictly out of court, his case may be more ambiguous than it first appears. Ryle certainly stresses the importance of behavioral dispositions and downplays the importance of phenomenal ones, sometimes even seeming to suggest that we could do without the latter entirely. Nevertheless, Ryle admits the relevance of such things as "silent colloquies" that others could not possibly overhear (1949, p. 184) and tunes in one's head consisting of "the ghosts of notes similar in all but loudness to the heard notes of the real tune" (1949, p. 269). For such reasons, Stuart Hampshire, one of Ryle's earliest critics and most careful readers, regards Ryle has having an "ambiguity of purpose" regarding the reduction of assertions about mental life entirely to statements about behavior (Hampshire 1950, p. 249). Despite his reputation, Ryle at times seems committed to the importance of internal, first-person phenomenology.

In light of this possibility, Ryle's short discussion of belief is interesting:

---

ascriptions.

> Certainly to believe that the ice is dangerously thin is to be unhesitant in telling oneself and others that it is thin, in acquiescing to other people's assertions to that effect, in objecting to statements of the contrary, in drawing consequences from the original proposition, and so forth. But it is also to be prone to skate warily, to shudder, to dwell in imagination on possible disasters and to warn other skaters. It is a propensity not only to make certain theoretical moves but also to make certain executive and imaginative moves as well as to have certain feelings (1949, p. 134-135).

If we set aside for a moment the standard picture of Ryle as bent on reducing all talk about mental life to talk about behavioral dispositions, this passage begins to look rather like an appeal to a *mix* of behavioral and phenomenal dispositions. Perhaps a bit optimistically, then, I would like to claim Ryle as the first (albeit wavering) advocate of *phenomenal dispositionalism* about belief.

2. Mixed Sets of Dispositions

The dispositional account of belief deals quite naturally with in-between cases of believing, cases in which it seems not quite appropriate to describe the subject as either fully believing or not believing the proposition in question. In this section, I provide a few examples of such mixed cases and sketch some of the patterns into which they tend to fall.[4] One of the central advantages I want to claim for the dispositional account is its facility in handling such cases.

Two Examples

Ellen studied Spanish for three years in high school. On the basis of her studies and her exposure to such Spanish words as 'mesa,' 'niña,' 'oreja,' and 'vaca,' she is willing, sincerely and cheerfully, to assent to the claim that all Spanish nouns ending in 'a' are feminine. Ellen has, however, occasionally come across certain words ending in 'ista,' such as 'anarquista' and 'bolchevista,' that can be used either as masculine or feminine (depending on the gender of the anarchist or bolshevik), and she uses them correctly as masculine when the situation demands. She would not assent to the claim that all Spanish nouns ending in 'a' are feminine if an 'ista' word came to mind as a counterexample; nevertheless, in most circumstances she would not recall such counterexamples.

[4] Stich (1983) is a good source of further cases, though Stich does not endorse a dispositional account of belief.

Does Ellen believe that all Spanish nouns ending in ‘a’ are feminine? Some of her dispositions accord with that belief; others do not. Whether it seems right to ascribe that belief to her varies contextually, depending on what dispositions interest us most. If we are considering which side she might take in a debate on the subject, it seems acceptable to say that she does believe that all Spanish nouns ending in ‘a’ are feminine. On the other hand, if we are interested in her skill as a speaker of Spanish and the likelihood of her making embarrassing gender errors in speech, it seems inappropriate to ascribe that belief to her. If we want to describe her cognitive state on the topic as carefully as possible, probably the best thing to do is to refuse to put the proposition “all Spanish nouns ending in ‘a’ are feminine” either simply in or simply out of some imaginary “belief box” in her head, and instead to sketch the mix of her dispositions as I have just done.

Geraldine’s teenage son Adam smokes marijuana. Usually Geraldine is unwilling to admit this to herself, and sometimes she adamantly denies it. Eating lunch with a friend, Geraldine can deplore her friend’s parenting because of his daughter’s drug problems while denying in all sincerity that Adam has any similar problems. Yet she feels afraid and suspicious when Adam slouches home late at night with bloodshot eyes, and when she accuses him of smoking pot, she sees through his denials. In a certain kind of mood, she would tell her therapist that she thinks Adam smokes marijuana, but in another kind of mood she would genuinely recant such a confession. When Geraldine’s husband voices concern on

the topic, Geraldine sincerely comes to her son's defense. What does Geraldine believe on the subject? Again, someone insisting on a simple "Yes, she believes he smokes marijuana" or "No, she doesn't" will be hard-pressed. Perhaps we could say that her beliefs on the subject change from situation to situation: When she is denying that her son smokes pot, she sincerely believes that he does not; when she is watching him creep in at 2:00 a.m., she sincerely believes that he does. But what does she believe *now,* while she's working intensely on a client's account and not giving the matter any thought? A simple yes-or-no answer seems misleading at best. Even if we want to describe her as self-deceived, she is at best only *partially* self-deceived, since there are conditions under which she would unhesitantly acknowledge that her son uses marijuana.

The cases of Ellen and Geraldine are not meant to depend on any lack of knowledge about their mental states, though lack of knowledge is a common source of hesitation in belief ascription. I do not want the reader to think I am putting forward an argument of the form: We cannot know what Ellen and Geraldine "really believe"; therefore, there is no fact about what they really believe. Rather, these examples are meant to be cases in which we *know* that the subject deviates partly from the stereotype for believing that *P*. I hope that, with these examples vividly before us, the reader will agree that in such cases, the person is in a state that cannot be quite accurately described as either simply believing or simply not believing that

P, and that a dispositional description of the subject's mental state adequately captures the facts.

Although some cases that are unmanageable on an all-or-nothing, Simple-Question view of belief become manageable simply upon recognition of *degrees* of belief, cases such as those described above do not yield to this approach. It is not that Ellen and Geraldine simply have a low degree of confidence (say .6 on a scale from 0 to 1) on the topic in question. Rather, they are disposed to feel in some situations quite confident in asserting one thing, while at the same time they are disposed to feel in other situations quite confident in asserting its opposite. The doxastic situation is far from the kind of steady uncertainty that one might feel, for example, about the outcome of a sporting event or the turning of a card. In light of this fact, it may be helpful to introduce some new terminology. The view of belief as simply an all-or-nothing matter we may call the *digital view;* the view of belief as always smoothly describable by particular degrees of confidence we may call the *analog view*. The cases on which I focus in this chapter are those unmanageable by either of these views. The dispositional account recommends handling these cases by describing in what ways the subject's dispositions conform to the stereotype for the belief in question and in what ways they deviate from it. Further questions may then be raised about the reasons for the match and mismatch of particular dispositions to the stereotype, opening avenues for both scientific research and everyday inquiry.

Normativity and Patterns of Deviation

The usefulness of classifying people's mental states by appeal to stereotypical dispositional patterns depends on the tendency of people to adhere to these patterns. If cases such as Ellen's and Geraldine's were the norm, the dispositional stereotypes of belief would have little purpose. As a general rule, however, people who conform to some parts of the stereotype are apt to conform to other parts also. Deviation from the stereotypes tends to fall into particular patterns as well, a few of which I will sketch briefly below.

The stereotypes capture more than merely statistical regularities, however. They capture something about how we think people *ought* to feel and behave. Something about Ellen's and Geraldine's phenomenology and behavior strikes us as normatively lacking, as incoherent or confused. We feel that if Ellen and Geraldine correctly reasoned things through, they wouldn't deviate from the dispositional stereotypes in the way they do. The conditional runs the other direction as well: Failures of reason will generally entrain failures to conform to the stereotypes.

This is not to say that conformity to all elements of the stereotypes is required by reason. For example, we can hardly convict someone of poor reasoning simply for not feeling disappointment upon suddenly learning that *P,* on which he had greatly counted, is false — strange though it may be in some

cases and contrary to the stereotype. At the same time, however, something about such cases leaves us uneasy. Our folk psychology and everyday dealings with other people are so thoroughly dependent on the accuracy of these stereotypes that perhaps there is a kind of *social accountability* to the stereotypes that pervades even those aspects of the stereotypes not shored up by the norms of reason. This, I think, is especially evident in the stereotypes associated with desires and personality traits, which are less thoroughly accountable to the strict demands of reason, and which consequently allow more room for social accountability to come undisguised into play. A person who is disposed greatly to enjoy ice cream on some occassions but to detest it on others, with no clear excusing conditions (such as detesting it only in times of grief), engenders a similar type of discomfort. We want to know whether, *really, deep down,* she likes ice cream or not. We want to fit her into our stereotypes, and there is some pressure on her actually to do so.[5]

Certain patterns of deviation, however, are pervasive enough that they don't at all strike us as strange, and in such cases we are much less likely to bring normative pressures to bear. A person's motor behavior and expectations might accord with a belief that *P,* but not most of her inward and outward verbal dispositions, as might be the case, for example, with a skier who always shifts his weight to the inside edge of the downhill ski *X*° through a turn but who could not tell anyone that this is what

[5] This topic is pursued in greater detail in Schwitzgebel and McGeer, "Psychological Dispositions: Revising the Philosophical Stereotype," unpublished MS.

he does. Alternately, people are often disposed to recognize and agree with assertions that *P* and able to answer correctly a question like *"P?* Yes or no?" yet not able to come up with *P* as an answer to a more open-ended question or to act upon the truth of *P* when uncued. My dispostions regarding the last names of many of my acquaintances from college follow this pattern. As a general rule, the more closely a mixed dispositional set matches a familiar pattern of deviation, the less puzzling it appears to us. At the other end of the spectrum would be cases in which the subject's dispositions regarding *P* vary widely in no recognizable pattern at all. In the extreme, we would have to describe such cases as insanity.

A careful account of such in-between cases will describe exactly in what respects the subject deviates from the stereotype of the belief in question and in what respects the subject accords with that stereotype (and, if relevant, with what degree of frequency such deviations will occur); it will look for a recognizable pattern in these deviations; and it will indicate which dispositions should count, in the present context, as the most important ones to the assessment. It may or may not have a normative element of the sort described in this subsection.

Deviation and Developmental Psychology

The dispositional account set forward in this chapter is especially useful for those interested in developmental psychology, since children, even more than adults, are apt to

have mixed dispositional states. Recall from chapter two Smith's daughter Zoë and her developing belief that her father is a philosopher (Smith 1982), or think of any belief ascription to a child where the concepts invoked do not much resemble the child's own. Although I argued in chapter two that belief ascription in such cases is often necessary and useful, the match to the dispositional stereotype is less than might be desirable. In such cases, as is true generally, whether a particular belief ascription is appropriate depends on the degree of match between the subject's dispositions and those dispositions in the stereotype that are important in the context.

The question of how well a child's dispositions match a given stereotype becomes even more difficult in discussing the general — one might say "theoretical" — beliefs of young children. Do three-year-olds, for example, think that beliefs can be false? (We might want to say that without this belief the child cannot have the concept of belief at all; see my treatment of this issue in chapter two.) As discussed in chapters two and four, there are respects in which their phenomenal and behavioral dispositions fail to accord with the stereotypes for this belief (Gopnik and Astington 1988; Wimmer and Perner 1983; Perner 1991b). At the same time, there are respects in which their behavior does accord with the stereotype. Researchers have found precocious behavior on after-the-fact explanatory tasks (e.g. "Why did she look under the piano instead of under the table?"; Wellman 1990) as well as when the experimenter conspires with the child to "trick" someone (Sullivan and Winner 1993; disputably

Hala, Chandler, and Fritz 1991). Although the preponderance of three-year-olds' dispositions do not seem to accord with the belief that beliefs can be false, it could be misleading simply to deny them this knowledge without qualification.

Similar examples abound. Piaget (1954) has argued, on the basis of reaching behavior, that five-month-old children do not believe objects continue to exist outside their perceptual fields, while Baillargeon (1987) and Spelke et al. (1992) have argued the contrary on the basis of the infant's looking behavior. (I will examine this case in more detail in the next chapter.) Or consider: At what age do children understand the past tense, given that their ability to use it is gradually acquired and generalized? In fact, every genuine case of Piagetian *décalage* — difference in timing between the development of skills tapping the same fundamental knowledge — can be described as a case of mixed dispositions regarding that fundamental knowledge.

A temptation arises in such cases to think that there must be a *moment* at which the child genuinely understands the facts in question and thus to think that apparent earlier expressions of the knowledge must be artifactual and that lapses afterward must be due to inaccessibility of the belief or "performance" (as opposed to "competence") difficulties. While skeptical inquiry into such potential shortcomings of developmental research is a *sine qua non* of good scientific method, it is unwarranted to insist adamantly that there *must* be such failures of methodology

when different tests point to development at different ages of capacities tapping the “same knowledge.” The latter insistence rests on the mistaken presupposition that such knowledge is unitary and acquired all at a moment rather than through a gradual, asynchronous shifting of a broad range of dispositions over a substantial span of development, as would seem on reflection to be the case, at least for the child’s most general, theoretical beliefs. The dispositional view of belief recommends a willingness to give up finding a simple answer to the question, Does the child *really* believe that such-and-such?

Talking about beliefs is scientifically useful because people with some of the dispositions in a stereotype will tend to have many of the other dispositions in that stereotype. Because of this, we can make generalizations and inductions on the basis of these stereotypes, and it is enormously convenient, even indispensible, to appeal to stereotypes in describing our mental lives. Still, when the match between stereotype and dispositional set *does* break down, as will often happen with young children and in cases of self-deception, in cases where things are not fully thought-through, and in many more cases besides, simple belief talk may no longer be appropriate, and appeals to the stereotype may have to be replaced with more complicated appeals to specific dispositions and sets of dispositions. And once the phenomenal and behavioral dispositions are made clear, it is a mistake to think there is

still some further question to be answered, namely, What does the subject *really* believe?

A Short List of Patterns of Deviation

It may be helpful to conclude this section by describing at least a few common patterns of deviation. This list is by no means exhaustive. How irrational the deviations on this list appear to us seems to be at least roughly proportional to the extent to which the subject could, by simple reflection, bring himself into line with the stereotype. Thus, "modularized believing" does not tend to strike us as particularly irrational, while "unreflective inconsistency" is more likely to strike us that way.

Modularized believing: It is common for a subject's dispositional profile to match that of the stereotype in a narrow area (or "domain") of expertise, but to deviate from the stereotype in most other domains and particularly with respect to the disposition to assent to *P* in inner speech. The example of the skier's knowledge of when to turn is meant to be an instance of this. In some cases, the dispositional profile can be brought into line with the stereotype by practice and reflection (see Karmiloff-Smith 1992), but often this will not be the case.

Unconscious beliefs: The history of psychoanalysis suggests that a subject may match a stereotype for believing that *P* in being disposed to claim that *P* under hypnosis or in free-association or in other of the techniques of psychoanalysis; and

the subject may exhibit hysterical or destructive symptoms that seem somehow consonant with a belief that *P,* though distorted; yet that subject may not be willing under normal circumstances to assent to *P,* even privately, because there is something unpleasant to the subject about the thought that *P* (see, e.g., Freud 1977). This idea has been generalized into the popular notion of the unconscious, according to which a person may be disposed to act in a variety of ways in accordance with the stereotype for believing (or desiring) that *P,* yet because of the unacceptability of the thought that *P,* not be disposed to admit to herself that *P* is the case. Different people may assess differently the frequency of such cases, though it seems hard to deny that they at least sometimes occur.

Self-deception: Cases classified by folk psychology in the category of "self-deception" may be a subset of cases of unconscious believing. Geraldine's attitude toward her teenage son may fit, imperfectly, into this category of deviation. In chapter seven, I will examine the case of self-deception in more detail.

Unreflective inconsistency: A subject may deviate from a stereotype simply because she fails to put two and two together. Ellen's case fits into this pattern. She matches the stereotype for believing that all Spanish nouns ending in 'a' are feminine in just those cases in which she is not reminded of a few exceptional nouns, and she deviates in cases in which those nouns become salient to her. We might suppose that with sufficient

reflection, Ellen would soon come to match fairly exactly the stereotype for believing that not all Spanish nouns ending in ‘a’ are feminine. In cases of this sort, one would expect a match to the stereotype for believing that *P* in just those cases in which the reasons against believing *P* are not salient.

Peripheral ignorance: Sometimes a person may fail to match a stereotype due to ignorance of related topics. Examples of this include the child who is uncertain about who the Pilgrims were and the case in which Joe believes there is Budweiser in the fridge but does not believe that Budweiser is a type of beer. Everyday intuition seems to be fairly competent at determining what the dispositional effects of any particular type of peripheral ignorance might be.

Developing beliefs: This type of deviation would seem to be closely related to the previous two. Acquiring a network of knowledge in a particular domain and forging that knowledge into the kind of coherent structure necessary to match consistently the stereotype for various beliefs in that domain necessarily takes a certain amount of time. During this period of transition the subject cannot be expected to match completely the stereotype for the developing belief. This position finds support in Vygotsky’s (1962, 1978) argument that children do not instantly acquire major new abilities and understandings, but rather must pass through a period during which they can exercise the knowledge or ability only with prompting or with proper structuring of the environment. As the child develops, less and

less of this external “scaffolding” is necessary for the child to meet with success, and the child passes to fully developed competency. In chapter seven I will examine two developmental cases in some detail.

Partial Forgetting: The process of forgetting or unlearning, in some ways the opposite of belief development, also does not take place all at once. I am in the midst, the reader will recall, of forgetting the last names of many of my college acquaintances. Some time ago, I could have rattled off their names easily; then it took more effort and sometimes the names did not come; now I can recall those names only with a prompt of some sort; perhaps later I will be able to pick them out in a forced-choice test; when I am eighty, I probably will not have any knowledge of them at all. The more demanding the recall situation and the fewer the prompts provided, the less likely someone in one of these intermediate stages of forgetting is to adhere to the stereotype of the belief that is being lost.

3. A Concern about Phenomenal Dispositionalism about Belief

Functionalists such as Lewis (1972, 1980) and Putnam (1966), as well as externalists about belief content such as Putnam (1975a), Burge (1979), and Davidson (1987), argue that the "content" of a belief is individuated not only in a forward-looking way, that is, by the phenomenology, behavior, and mental states it is apt to produce, but also at least in part in a backward-looking way, by how it came about (or at least how states of its type are apt to come about). In other words, both groups of philosophers highlight the importance of looking back at the causes of beliefs in determining their content. Won't the dispositionalist account run against the arguments invoked in favor of the backward-looking elements in these accounts?

Externalism and Phenomenal Dispositionalism

Externalists about belief hold that whether a subject believes that P, or whether the subject believes, instead, that Q, depends, at least sometimes, on facts about the world external to the subject herself. I will shortly describe an example. The dispositional account offered here is in fact compatible with our intuitions in the kinds of cases typically invoked to support externalism. In fact, the view comports more exactly with our intuitions in such cases than do the standard externalist views.

Consider Putnam's (1975a) example of Twin Earth, a planet identical to Earth in every respect except that where Earth has water, Twin Earth has twater, indistinguishable from water by any

of the tests available to inhabitants of Earth or Twin Earth but in fact a different chemical compound than H_2O. Wayne from Earth and Dwayne from from Twin Earth are molecule-for-molecule identical to each other (one might even suppose that Dwayne, through some freak occurrence, happens to be 90% genuine water). It seems intuitive to say that, despite the similarities between them, Wayne has beliefs about water, not twater, since that is what he interacts with on Earth, and Dwayne has beliefs about twater, not water (though both will, of course, use the word 'water' to describe what they see). If this is right, then it appears that the content of one's beliefs depends not only on what is in one's head, but also on one's environment and in particular on how one's beliefs were caused.

At first glance, it might seem that Wayne and Dwayne, being molecule-for-molecule identical to each other, could not possibly have different dispositions and thus must have the same beliefs on any dispositional account of belief. If this were so, then indeed the dispositional account of belief would run contrary to our intuitions in Twin-Earth-like cases. This would be unfortunate, perhaps, but not fatal: There is no guarantee that the most useful scientific or philosophical understandings of mind will accord with folk intuition in every respect.[6] As it turns out, however, dispositionalism about believing *is* compatible with such externalist intuitions, since dispositional

[6] Fodor's (1981) position of "methodological solipsism" (expanded from Putnam 1975a) is interesting in this respect, though he later revises it (1994). Roughly, it is the view that something like the folk concepts of belief, desire, etc. are appropriate for psychological theorizing about the mind, but these concepts must be purged of any of their externalist consequences.

properties themselves may be defined in part “externally,” i.e., with reference to the organism’s past or its environment. Only Wayne has the disposition to regard a present instance of water as an instance of the same kind of stuff Wayne drank as a child. Only Dwayne has the disposition to use the word ‘water’ intending to refer to the same kind of stuff people in his community on Twin Earth refer to by using that word. Someone who believes that the meaning of a sentence is in part determined by factors external to the individual uttering those sentences has an additional pool of externally individuated dispositions to draw from in distinguishing Wayne from Dwayne. When Wayne utters the sentence ‘water is clear and potable,’ he is uttering a sentence that means water is clear and potable; when Dwayne makes exactly the same sounds, he is uttering a sentence that means *twater* is clear and potable. Thus, if sentence meaning is in part determined by external factors, Wayne will be disposed to say one kind of thing, while Dwayne will be disposed to say quite another.

So there are at least *some* dispositions Wayne and Dwayne do not share. The question about whether we should describe them as having the same belief, then, depends on whether these differences are regarded as important enough in the context of ascription to warrant differential treatment of Wayne and Dwayne. If one chooses to focus on utterance meanings, and if these are individuated externally, or if one focuses on dispositions invoking one’s past or one’s community, one can fairly readily be

drawn into regarding the two men as having different beliefs. If one focuses instead on what it is like from the inside, on phenomenology and motoric behavior, and especially if one is not an externalist about linguistic meaning, one may find oneself drawn in the internalist direction.[7] An internalist dispositionalist would hold that externally individuated dispositions are *never* relevant, for the purposes of philosophy or science, to the assessment of belief.

Although Putnam makes a good case for the intuitiveness of describing Wayne's and Dwayne's beliefs differently (similarly for Burge and Davidson with respect to their examples), in some contexts the intuitions are not so clear. For instance, let us suppose that Wayne and Dwayne are both environmental engineers working on a large water-treatment project. Miraculously, Wayne and Dwayne are teleported to each other's worlds. Wayne's coworkers may be concerned about Dwayne's ability to continue with the project. Doesn't it seem right to say that they shouldn't worry because Dwayne's beliefs on the processes of water treatment are exactly the same as Wayne's?

Given that our intuitions on the Twin Earth and the other externalist cases are somewhat ambivalent and context dependent, as I think they are, then the dispositional account of belief I have offered has an advantage over standard externalist accounts of such cases, since it provides room for such ambivalence and allows us to predict contexts in which the intuitions may go one

[7] Dretske (1995), however, argues that even phenomenal experiences should be individuated externally.

direction or another. In the water-treatment case, the dispositions Wayne and Dwayne *do* share are the focus of concern, and so the dispositionalist account would predict an inclination to regard the two as having the same belief. In another case, perhaps where we are particularly concerned with what kind of stuff Wayne and Dwayne intend to pick out by means of their word 'water,' the dispositionalist account may pull in the externalist direction (depending on whether you think Wayne's and Dwayne's words *do* refer to different kinds of stuff). Whereas the dispositionalist account can accommodate intuitions pulling in both directions and to some extent predict on the basis of context in which direction our intuitions will be pulled, standard externalist accounts must stand fast with an unchangeable answer: that what Wayne and Dwayne believe really *is* different; thus externalists are forced to try to explain away internalist intuitions the dispositionalist account handles quite naturally.

Functionalism and Phenomenal Dispositionalism

What about functionalist arguments for the necessity of invoking backward-looking as well as forward-looking criteria for belief individuation? Functionalists hold that what makes a state a belief is its causal role in the system in which it takes a part, or the causal role that states of its type typically play in systems of the type in which it takes a part (Lewis 1980; Shoemaker 1981; Block 1978). A state's causal role has both

forward-looking and backward-looking components — it is both apt to be caused by certain kinds of events and apt to cause certain kinds of events. Pain is the favorite example: It is apt to be produced by, among other things, pinchings, pokings, fire, pressure, and bodily injury, and it is apt to produce, in turn, groaning, writhing, disrupted thoughts, and avoidance. Although it is common for functionalists considering the individuation of mental states to argue for the importance of causal role generally, it is not as common to find arguments for the importance of including the backward-looking elements of causal role as opposed to including only at the forward-looking elements.

Shoemaker is an exception. He begins his 1981 paper with an attack on behaviorism like Chisholm's (1957) attack discussed above: Because how one's beliefs dispose one to behave depends on one's desires and how one's desires dispose one to behave depends on one's beliefs, it will be impossible to reduce talk about mental states to any other kind of talk so long as one appeals only to behavioral dispositions. Shoemaker, however, *does* take as his aim the redefinition of mental predicates in terms of predicates containing no mental predicates. Shoemaker says,

> Let us say that a state (mental or otherwise) is functionally definable in the strong sense just in case it is expressible by a functional predicate that contains no mental predicates (or mental terminology) whatever.... It is functional states in this sense which functionalism takes mental states to be (1981, p. 95).

So long as one's task is to provide for mental states functional definitions in this strong sense, post-Rylean, anti-behaviorist

arguments like Chisholm's show that mere appeal to forward-looking dispositions will not do. Functionalists appeal, therefore, not only to dispositions to behave but also to the typical physical causes of mental states and also to the causal relations between mental states, on the understanding that the whole bundle of mental states, taken together, can in principle be characterized wholly in terms of physically (or at least non-mentally) described inputs and outputs (Lewis 1972; Block 1978). Since it is not part of the project of phenomenal dispositionalism to characterize mental predicates by means of non-mental predicates, the functionalist's reasons for wanting to appeal to the backward-looking relations of mental states do not apply.

Perhaps, however, there is some warrant for a revised functionalism that characterizes and individuates mental states both dispositionally *and* in terms of how they are apt to come about, but at the same time does not require that mental predicates be in-principle characterizable by non-mental ones — a functionalist account, in other words, that does not treat phenomenology simply as falling out of the functional relations but rather treats phenomenology as itself one of the relata. I have no serious objections to such a view, although in the case of belief in particular I am inclined to make the stronger claim that once one takes phenomenal dispositions seriously, an adequate characterization of what it is for a subject to believe something does not require appeal beyond the dispositional

features of the subject’s mental life. To argue otherwise would require quite a different set of objections than can readily be drawn from the functionalist literature.

4. Beliefs, Causation, and Explanation

Joe rises off the couch and heads for the fridge. Intuitively, we explain this behavior by appealing to various mental states of his: He feels thirsty. He wants a beer. He thinks that there is a beer in the fridge. Moreover, we hold that these mental states are causally effective in getting him to the fridge. In general, it is supposed, mental states like belief both *cause* and *explain* much of our behavior.

Many philosophers of mind today accept something like this intuitive picture. Thus, for example, Fodor regards it as an essential feature of mental states like belief that they cause behavior and can be invoked to explain it (1987, p. 12-14). One of the primary tasks of Dretske's 1988 book is to show how states with indicative content, like beliefs (see above, chapter four), can cause and explain behavior. Searle (1984) also argues that beliefs play a crucial role in causing and explaining behavior.

I accept this picture of belief, although I would hasten to add that beliefs cause and explain phenomenology (and other internal changes) as well as behavior. Nevertheless, several people have objected that a dispositional account of belief leaves no room for belief to play such a causal and explanatory role.[8] If believing just is being disposed towards certain behavior and phenomenology, the objection goes, it is illegitimate to say that beliefs cause or explain that behavior and phenomenology. The objection has even more bite if we take the explanandum to be itself a disposition. It seems natural,

for example, to explain the disposition to assent under certain circumstances to utterances of the form "P?" by appealing to the fact that the subject believes that P. But if believing that P just is a matter of having such dispositions, then seemingly the belief cannot be invoked to explain the presence of those same dispositions.

I will break my response to this objection into several parts. First, let us consider the question of whether a belief, regarded as a disposition to manifest certain phenomenology and behavior, should be thought of as causing that phenomenology and behavior when it is manifested. If a negative answer is urged to this question, presumably it is done so on the basis of a *general* commitment to the position that dispositional states do not cause their manifestations. Consider, then, the general question of whether dispositions can cause their manifestations. For concreteness, consider the case of solubility. (Solubility is indisputably regarded as dispositional: Something is soluble in water just in case it is disposed, under normal conditions, to dissolve in water.) Is something's solubility in water (the disposition) a cause of its dissolving when placed in water (the manifestation)?

Philosophers interested in the metaphysics of dispositions are, in fact, divided on the question of whether dispositions cause their manifestations. David Armstrong (1968, 1969) and William Rozeboom (1978) have argued that dispositions do cause their manifestations. They argue for the point in essentially

[8] This point has been put to me most vividly by Max Deutsch and John Searle.

the same way, although Rozeboom adds several complications absent in Armstrong. The argument runs like this. For every dispositional property, there must be some *categorical basis* -- i.e., some non-dispositional property causally responsible for the dispositional manifestation when the triggering conditions are met. But, in fact, the dispositional property is nothing over and above its categorical basis; indeed, it is to be identified with it. Since categorical bases, by stipulation, cause dispositional manifestations, so also do dispositions. On this view, then, beliefs regarded as dispositions can cause their phenomenal and behavioral manifestations, and one version of the objection mounted two paragraphs back is defeated.

Another view of dispositions denies the existence of categorical bases for dispositions. Ryle (1949) is typically read as holding such a view (e.g., by Armstrong 1968; Mackie 1973; Prior 1985). A proponent of this view regards claims about dispositional properties as bare conditional claims, asserting a connection between trigger and manifestation, but requiring no commitment to the existence of an underlying property responsible for the maintenance of that connection. On this view, it would appear that dispositions do not cause their manifestations. If a disposition is simply a regularity or the obtaining of a conditional fact, it cannot be a cause, for although regularities and conditional facts may suggest the existence of causal relations, it seems that they are not the right sort of things themselves to be causes.

Armstrong (1969) and Elizabeth Prior (1985) have argued against the Rylean view, contending that it flies in the face of the common intuition that there must be something in the world that makes the dispositional claims true, some persisting feature of the object to which the dispositional property is ascribed that causes the manifestation when the triggering condition is met. I accept their argument on this point. In any case, the old Rylean view of dispositions without bases has something of a verificationist feel that sits at best uncomfortably with the realist talk of beliefs as causes of behavior that is presupposed by the objection I am addressing. After all, if dispositions can be manifested without the existence of some underlying cause in the object that has the dispositional property, then presumably human behavioral and phenomenological dispositions can operate the same way; and if they can, then the case for the existence of beliefs as causes of such behavior and phenomenology is on shaky ground. Either such behavioral and phenomenal dispositions have no categorical basis, in which case we ought not think that they are the causal result of some belief, or they do have a categorical basis, in which case the Rylean approach to these dispositions is out.

A third view of dispositions grants the existence of categorical bases for dispositions, but refuses to equate dispositions with those bases. Prior (1985), for example, advocates “functionalism” about dispositions, on which a dispositional property is a higher-order property -- the property of having one or another non-dispositional property, or basis,

that plays the causal role of producing the manifestation when the triggering conditions are met. On Prior's view, the categorical basis for any disposition is a sufficient cause of the manifestation, given the triggering condition, and therefore the dispositions themselves cannot cause their manifestations: There is no causal work left over for them to do, once the basis has done its business. So, for example, something about the ionic structure of salt causes it to dissolve when placed in water. That something is the categorical basis of its dissolving. The property of having some structure, ionic or otherwise, that results in dissolution when placed in water is the property of being disposed to dissolve in water. But this property does not cause the dissolution; rather the ionic structure of the salt does.

Note that neither on Armstrong's and Rozeboom's nor on Prior's view does having the categorical basis cause an object to have the dispositional property: Having the categorical basis causes the dispositional *manifestation* in the relevant circumstances. Having the catergorical basis is either identified with having the dipositional property (Armstrong, Rozeboom) or having some basis or other of the right sort is identified with having the dispositional property (Prior).

I have no particular quarrel with either view of dispositions. But, if I accept Prior's view, does my view of belief then imply that beliefs cannot cause the behavior and phenomenology belonging to their dispositional stereotypes, since

dispositions on Prior's view do not cause their manifestations? It does, if beliefs are themselves seen as complicated dispositions, consisting of a conjunction of the individual dispositions in their stereotypes. The view I espouse, however, is not committed to treating belief in that way. So long as there is a categorical or causal basis for the phenomenology and behavior in question, the belief can be identified with that basis, regardless of whether dispositions themselves are so identified.

Let me clarify this point just a bit. In the first section of this chapter, I offered an account of what it is to *believe* something but no account of what a *belief* is. I do not think an account of the latter sort as useful as the former, in part because thinking too much in terms of beliefs and too little in terms of believing strengthens the container metaphor for belief, repudiated in chapter five. After all, *beliefs* seem to be things *in* the head (or at least locatable somewhere). Nevertheless, it is necessary from time to time to talk about beliefs, and so a good account of them is necessary. Here, then, is my idea: A *belief* is a state of a creature causally responsible for its responding in ways that match the appropriate dispositional stereotype.[9] *Having* a belief, then, is being in such a state (and in a causally rich world, as I suppose ours to be, anyone who believes that P -- i.e., anyone who matches to an appropriate degree and in appropriate respects the dispositional stereotype

[9] One might want to add further conditions to this definition, if that be thought necessary to get at the right part of the causal chain.

for believing that P -- will also have the belief that P). It is then trivially true that beliefs cause phenomenology and behavior.

So far I have talked a lot about causation and not at all about explanation, but the objection requires that the dispositional account allow not only for beliefs to *cause* phenomenology and behavior but also for beliefs to *explain* phenomenology and behavior. However, once we allow that beliefs cause phenomenology and behavior, it is a quick step to the conclusion they can be invoked to explain it. David Lewis (1986a; similarly, Humphreys 1989) argues that to explain an event simply is to cite information about its causal history. On this account of explanation, surely, beliefs can explain behavior. But even on accounts of explanation that do not equate explanation with providing causal information, paradigmatic explanations of events cite the causes of those events. Why did the water boil? Because the stove was turned on. Even the appearance of ‘cause’ in ‘because’ suggests this connection between causes and explanations. If we explain why the child tripped by citing (a.) the rock’s being in the trajectory of his foot and (b.) his not paying attention to where he was going, we have given a partially physical and a partially mental explanation of the event; and in both cases what we have done is cite causes.

I hope that I have dealt adequately with the objector’s concern about the ability of beliefs, on my account, to cause and

explain phenomenology and behavior. I will now tackle the question of the causation and explanation of particular dispositions within the stereotype, beginning with the issue of explanation. It is important here to keep clear in one's mind the difference between the explanation of particular *dispositional manifestations* and the explanation of particular *dispositions*. My response to the first version of the objection turned on treating beliefs as the bases that cause, and thereby explain, their behavioral and phenomenal manifestations. We are now turning our attention to the question of whether beliefs, on my account, can explain the presence of particular dispositions. A similar response is not open to the this version of the objection: Categorical bases do not cause the dispositions for which they are the bases.

Intuitively, it seems plausible to say that Joe's believing that there is beer in the fridge explains his disposition to assent to the claim that there is beer in the fridge (*ceteris paribus*). The supposition of the objector is that we would have to reject this intuition on the dispositional account of belief: If to believe that P is simply to have a variety of dispositions of this sort, believing that P cannot explain the presence of those very dispositions.

Let me sort out what is right and what is wrong in this objection. Certainly we cannot explain the tendency of salt to dissolve in water by appealing to its disposition to dissolve in water; nor can we explain the presence of the entire range of

dispositions in the stereotype for a belief by appealing to the existence of that belief. However, it does seem intuitive to say that we can explain the tendency of salt to dissolve in holy water by appealing to its tendency to dissolve in water in general. This case is in important respects parallel to explaining Joe's disposition to assent by appeal to his belief. It is intuitively acceptable to explain the presence of one disposition by appealing to a larger set of dispositions that encompasses it.

Consider, as a similar case, Kepler's laws of planetary motion. Although these laws predict the position of the planets with substantial accuracy, they do not (by themselves) reveal any cause of the motions or in any way add to our knowledge of the planets, except in so far as they reveal a pattern in the planets' motions that had not before been noticed. Nevertheless, it seems right to say that we can explain the appearance of a planet in one part or another of the night sky by appealing to Kepler's laws. Fitting the planet's motions into an easily comprehensible pattern of regularities is a way of explaining it. The planet was at such-and-such a place three weeks ago, so according to these equations governing its regular motion, it ought to be in this place now. Even Newtonian mechanics might be thought to explain in the same way. Explanations of this sort work by fitting isolated facts or events into a larger pattern, even when no explanation is available as to why that pattern is one way rather than another. Similarly, then, one can also explain particular behavioral and phenomenal dispositions by

fitting them into the larger dispositional stereotypes of belief. So again, the objection fails.

Perhaps, however, it will seem necessary to offer an account of belief on which the presence of the belief is *causally* responsible for the individual dispositions in the stereotype and on which the *whole* pattern of those dispositions is to be explained by appeal to the presence of that belief. Here, finally, we have a pair of demands that the dispositional account cannot satisfy.

These demands do not have the intuitive appeal of the demands with which the dispositional account is compatible. While most of us would find it intuitive to say that Joe's belief causes and explains his trip to the fridge, and even that it explains his disposition to assent to certain statements, it is not equally intuitive to say that Joe's belief *causes* his disposition to assent to certain statements; nor is it very intuitive to say that Joe's belief explains the presence, not of each disposition considered individually, but of the entire range of the dispositions in the stereotype, considered as a whole. Even if we did have these intuitions, I see no reason to regard them as inviolate in the face of an otherwise appealing account of belief that contravenes them.

I believe there are also good independent reasons to reject these particular intuitions. If believing *causes* one to have all the dispositions in the stereotype associated with that belief (and thereby explains the match to that stereotype), then

believing must be a state distinct from matching the dispositional stereotype for P. When two states are not distinct, one cannot cause the other, just as something's being three-angled cannot cause it to be three-sided or something's being an election in 1996 cannot cause it to be an election full stop. (Those who hold that a disposition causes its manifestation hold that the disposition is distinct from its manifestation; the categorical basis, however, not being distinct from the disposition cannot cause it, as described above.) But surely it is fanciful to think that there is some distinct state of the mind, separate from having the range of dispositions in the stereotype for believing that P, that is the state of believing that P. How could we identify such a state, apart from appealing to the dispositions it is apt to produce? And what great benefit would there be in talking about such a state? Even if we supposed such a state to exist, I cannot but think that it would be more profitable to talk about a creature's overall dispositional make-up, and tie believing to that, than to single out such an elusive ghost as the proper referent of such an important word as 'belief.'

5. Conclusion

In this chapter, I have set out a novel account of belief. Like Ryle, I suggest that having a belief is nothing more or less than having a certain range of dispositions. Unlike Ryle, however, I emphasize the *phenomenal* dispositions involved in believing and see no reason to downplay or be reductivist regarding talk about our internal mental lives. I also go beyond Ryle in introducing the notion of a *dispositional stereotype* against which a person's dispositional profile can be matched, to help make sense of and provide a structure for talking about cases of what I have called *in-between* believing. I discussed some cases of in-between believing in more detail and outlined some common patterns of deviation from the dispositional stereotypes for belief. Finally, I addressed some concerns about the dispositional account that might naturally arise out of an externalist or functionalist view of belief or out of attention to issues of explanation and causation. I will close by addressing the question of how compatible my account is with the idea that beliefs are real, concrete states of the brain, discernible and classifiable, at least potentially, to an advanced science with substantial knowledge about how the brain works.

The relation between this view and my account of belief is perhaps best approached with the help of an analogy. I ask the reader to imagine a nineteenth-century understanding of disease before the advent of the germ theory. We will not imagine it as

the messy thing it actually was, but instead in a rather idealized fashion. To have a disease, on the empiricist view I am imagining, is simply to have some cluster of symptoms. These symptoms tend to cluster together into general patterns, and we may label these patterns of symptoms with different names: dropsy, diphtheria, tuberculosis, etc. In diagnosing a patient, one examines that patient's symptoms and determines which of these named clusters she most closely approximates. (We will ignore the little complication of discovering new diseases.) The more closely a patient's symptoms match the cluster of symptoms associated with a certain disease, the more appropriate it is to describe the patient as having that disease. A patient whose symptoms deviate from all the known stereotypes of disease cannot be said simply to have one disease or another; to describe that patient's condition accurately, one can only give a list of particular symptoms.

Those holding this model of disease would know, of course, that there must be some set of causes for the tendency of symptoms to cluster together and for the clustering together of particular symptoms in particular cases. However, since they admit ignorance regarding what exactly these causes might be, they must make do with an account of disease that appeals only the patient's match to a stereotypical profile of symptoms. It may or it may not turn out that there is a single, simple cause, such as the possession of one single physical characteristic (e.g., infestation by a certain type of microbe the immune system cannot effectively suppress), at the root of any particular

clustering of symptoms. If it did turn out this way, then a restructuring of the understanding of disease would probably be desirable, and in the process of such a restructuring it may begin to look more like a simple yes-or-no question (or a simple analog matter of degree) whether a person has a disease. On the other hand, it may turn out that diseases in fact have no such simple causes, that symptoms are clustered together for reasons too complicated for us to reduce to a single, labeled cause, and the symptom-cluster account of disease is the best account available to human understanding. The pre-germ account of disease is justified in either case, since nothing better is to be had for the time being, despite the fact that it is reasonable to suppose that it may be replaced.

I would suggest that we are in a similar position with regard to beliefs. It may, or it may not, turn out that there are some fairly straightforward and scientifically scrutible bodily causes for the clustering together of dispositions into the stereotypes with which we are familiar. If this does turn out to be the case — if beliefs really are strongly concrete and observable in this way — then we may wish to restructure our understanding of belief around these causes. But until such causes are discovered, if ever they are, a symptom-based account of belief is fully warranted. Embrace, therefore, as robust and optimistic a realism about belief as you wish: It is not incompatible with accepting, at least for the time being, the dispositional account of belief offered here.

Chapter Seven

Applications of the Account

In chapter five I declared my intention to develop an account of belief that has practical utility for working philosophers and psychologists. To have practical utility, an account must promote clear thinking on the topic at hand, it must help its users make sense of current research, and it must direct their attention away from fruitless inquiries into more productive ones. I believe that the account presented in the previous chapter has this kind of practical utility. The reader has seen the utility of the account in handling the many examples of "in-between" believing presented in that chapter; but to see the real value of the account for philosophical and psychological research, it is necessary to see how the account interfaces with actual contemporary research in these fields.

In this chapter, I will apply my dispositional account of belief to four areas of current research, two in philosophy and two in developmental psychology. We will see philosophers and psychologists repeatedly stumble over the kinds of in-between cases of belief that have been the focus of my attention in these chapters. And we will see energy directed away from useful avenues of inquiry into counterproductive attempts to squeeze genuinely mixed cases of believing into simple all-or-nothing descriptions.

1. Two Philosophical Puzzles

I will begin by describing two philosophical puzzles into which I think we can gain insight by application of my account. I will then show how my account applies to these puzzles and other potentially troublesome similar cases.

Kripke's Puzzle about Belief

The first of the two philosophical puzzles I will be discussing here is put forward in Saul Kripke's (1979) paper, "A Puzzle about Belief". In this paper, Kripke describes several cases in which he thinks standard assumptions about belief lead to paradox. The most fully fleshed-out of these problem cases is that of Pierre, a native French speaker who does not know English, but who grows up reading travel guides and hearing tales of the beauty and magnificence of a certain distant town called 'Londres'. If someone were to ask Pierre, in French, whether he thought that town was pretty, he would assent, and it seems quite natural to say that he believes that London is pretty. Later in his life, Pierre moves to London without knowing it is the same town he calls 'Londres', and he thinks it an ugly place. He would heartily assent to the English sentence, 'London is not pretty'. At the same time, since he has not learned that 'Londres' is the French word for 'London', he would still be willing to claim, in French, that 'Londres est jolie'. He thinks, in other words, that 'Londres' and 'London' name different places, the first pretty and the second not.

Now, Kripke argues, we are on the edge of paradox. If we take Pierre's French utterances seriously, we seem to be compelled to say that Pierre thinks that London is pretty, 'London is pretty' being the English translation of the French sentence to which Pierre sincerely assents. On the other hand, if we take Pierre's English utterances seriously, we seem compelled to say that he thinks that London is not pretty. So Pierre would appear to have contradictory beliefs. Even, however, if we are comfortable describing people as having contradictory beliefs in some cases, in Pierre's case the matter is especially strange: He would seem to be guilty of no logical error but simply a lack of information. It seems unfair to convict him of logical inconsistency.

Can we escape the difficulty by denying either (a.) that Pierre believes that London is not pretty or (b.) that Pierre believes that London is pretty? Rejecting the first claim seems pretty much out of the question: Pierre lives in London and sincerely says that it is not pretty. Rejecting the second claim is a little more tempting. Perhaps Pierre no longer believes that London is pretty. Certainly he did once believe this. He and his French buddies dreamed of someday visiting the beautiful town they called 'Londres' and read about in travel books. But if he did once believe that London is pretty, then ought we not allow that he still believes it? He will still assent to all the same claims, expressed in French, to which he would have assented as a youth. If he ran into his old French buddies, they would see in his eyes not disgust but the familiar dreamy glaze as he

talked about someday visiting the beautiful town of ‘Londres’. If everything he ever learned in England were rubbed from his brain, the memories and opinions he still has from France would be amply sufficient to ascribe him the belief that London is pretty.

Note that if we take Pierre’s English utterances seriously and say Pierre does not believe that London is pretty and then turn around and say that Pierre does believe that London is pretty on the basis of his French utterances, it is not only Pierre who has contradictory beliefs, but we ourselves.

So what does Pierre really believe about London? Does he really believe it is pretty, or does he really believe it isn’t? Or can we make sense of the claim that Pierre really believes both? Or does he, perhaps, have *no* beliefs about London’s beauty? In the face of apparently decisive objections to all these options, Kripke announces that the puzzle here is a genuine puzzle, on a par with such famous philosophical puzzles as the Liar’s Paradox.

A small body of literature has grown up in response to Kripke’s puzzle. Richard Garrett (1991), elaborating on an earlier suggestion by Hilary Putnam (1979), argues that Kripke’s puzzle shows that all our beliefs about any object must be qualified by identifying knowledge that allows us to uniquely single out that object. We should not say that Pierre has any bare, unqualified beliefs simply *about London*. Rather, Pierre believes that London, identified in whatever way he associates

with the name ‘Londres’, is pretty; and he believes that London, identified in whatever way he associates with the name ‘London’, is not pretty. So long as his associations with the name ‘Londres’ are not the same as his associations with the name ‘London’, Pierre’s beliefs are not contradictory, and the puzzle disappears.

Appealing as this solution might seem, it has difficulties. First, we should note that it is one thing to say that in order to believe anything about London we must have some identifying knowledge of it; it is quite another thing to claim, as Garrett does, that this knowledge is implicit in and qualifies all our other beliefs about the city. You and I may both believe a lot of things about London that don’t uniquely identify it -- such as that it is a big city in England with red double-decker buses and good Indian food -- but if I identify London as the largest city in England and you identify it as the capitol of England, none of my beliefs can, on Garrett’s view, be either the same as or inconsistent with any of yours. If I claim that London is pretty and you claim that it is not pretty, we have not, despite appearances, contradicted each other. Since each statement is qualified by different identifying knowledge, neither statement, by Garrett’s own assertion, entails the denial of the other. Surely this is a rather counterintuitive position to endorse for the sake of escaping Kripke’s puzzle. Yet we must endorse it, if Garrett’s solution is to work, for it is the very fact that Pierre’s two beliefs about London do *not* contradict each other,

each being qualified by a different identification of that town, that Garrett explicitly leans on to justify ascribing both beliefs to Pierre.

Robert Fogelin (1994) proposes a rather different solution to the puzzle. Fogelin argues that we should see Pierre as having what he calls a "divided belief system". Pierre's beliefs, on Fogelin's view, are divisible into two distinct subsystems, a Francophone system and an Anglophone system. Pierre's Francophone system subscribes to the belief that London is pretty; Pierre's Anglophone system subscribes to the contradictory belief. It is a mistake, on Fogelin's view, to insist on answering the question whether Pierre, considered as a whole, believes that London is pretty; we can only answer the question when it is relativized to one or the other of Pierre's two subsystems.

Some difficulties also arise for this approach to Kripke's puzzle. First, it seems to make Pierre's problem a problem of self-knowledge. If Francophone Pierre could only gain access to the beliefs of Anglophone Pierre, then perhaps he could spot the inconsistency between the two systems and make some efforts to repair it. But surely this description mistakes the case: No amount of introspective prowess can get Pierre out of his situation. What he is lacking is not some piece of knowledge about *himself,* but rather a piece of knowledge about the coreferentiality of the words 'London' and 'Londres'.

Still more troubling, to my mind, is the plethora of issues that arise about the mechanics of Fogelin's division of the mind. Francophone Pierre and Anglophone Pierre presumably share many beliefs, even if they do not agree about the aesthetic merits of London. Are these beliefs somehow encoded twice in Pierre's brain, once in English and once in French, or are they only encoded once, with the Francophone system and the Anglophone system equally capable of accessing most of them? If they are encoded twice, that seems like an awful waste of resources. If there is one common pool of beliefs to which both systems have access, how is it that beliefs, one way or the other, about London's beauty came to be excluded from that pool? What is the mechanism that separates Pierre's two subsystems of belief, and to what extent is communication possible between the parts? Fogelin also suggests other ways of dividing the mind -- for instance, a person might have beliefs in a subsystem of his mind activated when he is drunk that he does not have in the subsystem that is active when he is sober. One might ask whether different divisions of the mind can cross-cut each other; if so, can they act as a bridge for communication between those parts they cross-cut?

I put forward these questions to bring out the serious nature of the claim that the mind is divided into subsystems; claims of this sort, if they are to be taken literally, raise a variety of issues. It makes sense to consider such issues about, for example, the division of the visual system from the rest of the brain. The anatomical, neurophysiological, and cognitive

evidence for such a division is strong, and we do want to know what the mechanisms of isolation and communication are, how and whether the division cross-cuts other plausible divisions in the mind, and to what extent information must be re-encoded within different systems. It seems a radical step to say that Pierre is similarly *literally* divided into Francophone and Anglophone belief subsystems; but if the division is merely a metaphorical one, it's hard to see how it will do the necessary work.

Most people's first reaction to Kripke's puzzle is that its solution must be easy. And I do think a proper solution, which falls out of the account of belief offered in the previous chapter, has something of an easy feel about it. On the other hand, the variety and complexity of the solutions that have been offered to this puzzle belies the hunch that the problem is a cinch; we should not underplay the difficulty of Kripke's puzzle.

Self-Deception

The second philosophical puzzle I will consider is the case of self-deception. The philosophical literature on self-deception, like the literature on Kripke's puzzle, presents situations in which it is difficult to say whether a particular belief ascription is appropriate or not. Such a case is described by Amelie Rorty in a 1988 paper on the topic:

> If anyone is ever self-deceived, Dr. Laetitia Androvna is that person. A specialist in the diagnosis of cancer, whose fascination for the obscure does not usually blind her to the obvious, she has begun to misdescribe and ignore symptoms that the most junior premedical student would recognize as the unmistakable symptoms of the late

> stages of a currently incurable form of cancer. Normally introspective, given to consulting friends on important matters, she now uncharacteristically deflects their questions and attempts to discuss her condition. Nevertheless, also uncharacteristically, she is bringing her practical and financial affairs into order: Though young and by no means affluent, she is drawing up a detailed will. Never a serious correspondent, reticent about matters of affection, she has taken to writing effusive letters to distant friends and relatives, intimating farewells, and urging them to visit her soon. Let us suppose that none of this behavior is deliberately deceptive: She has not adopted a policy of stoic silence to spare her friends. On the surface of it, as far as she knows, she is hiding nothing (1988, p. 11).

Let us now consider the following question: Does Androvna believe that she has cancer? Different facts about Androvna seem to point in different directions. On the one hand, Androvna's drawing up her will and writing effusive letters are actions that seem inexplicable unless they arise somehow from the belief that she has cancer. On the other hand, Androvna sincerely and consistently disavows having this belief, argues that the evidence for cancer is inconclusive, thinks her brother rude and ignorant when he suggests that she may have cancer, and so forth. *These* actions seem difficult to explain unless we say that Androvna does *not* believe that she has cancer. We would appear to have, then, a dilemma: Say that Androvna does believe she has cancer and one subset of her actions becomes inexplicable; say that she doesn't believe it and a different subset of her actions becomes inexplicable. Our everyday intuitions about belief ascription don't weigh in strongly in favor of one option or the other. The phrase 'self-deception' seems to suggest that she has somehow managed to fool herself into believing that she doesn't

have cancer, and therefore doesn't believe that she has cancer. On the other hand, it seems equally natural to say that certain actions reveal that "deep down" she *does* believe that she has cancer.

Some authors, such as Rorty (1972, 1988), David Pears (1984) and perhaps Donald Davidson (1982a, 1985b), have attempted to escape this dilemma by pursuing an alternative similar to Fogelin's proposal for dealing with Kripke's puzzle: They have suggested splintering the self into discrete subsystems, each with only partial access to the other's cognitions. Once this is done, the option is open to say that Androvna has one subsystem that believes that she has cancer and another subsystem that does not. The actions that seem to require the belief that she has cancer are actions that are directed by, or somehow informed especially by, the subsystem that has that belief. The actions that seem to require absence of this belief are those directed or informed by the other subsystem. A variant of this strategy, advocated by Raphael Demos (1960) and Brian McLaughlin (1988), does not strictly insist on dividing the mind into subsystems but rather allows the unpleasant belief (in this case Androvna's belief that she has cancer) to retreat, in some range of circumstances, into "inaccessibility" while the contrary belief (that she does not have cancer) is held in some more accessible fashion.

Other authors, such as Robert Audi (1982, 1985) and Kent Bach (1981), have argued that the self-deceiver really, genuinely

believes only the unpleasant proposition, and not the more desirable one. Actions that seem to depend on *not* having the unpleasant belief are then explained as the effect of suppressing the belief or acting on the basis of persistent avowals to the contrary. Still others, such as Alfred Mele (1987a), have argued the opposite: What the subject really believes is rather the more pleasant proposition -- Androvna really believes that she does not have cancer. This belief emerges as the product of various biasing strategies, such as weighing evidence in favor of the preferred belief more heavily than it warrants or only making an effort to gather evidence on one side of the issue. If one occasionally acts, as Androvna does, on the basis of the unpleasant truth, doing so must be the product of a momentary lapse in one's ordinarily more pleasant convictions.

Each of these approaches to self-deception has some plausibility, and it is difficult to find a firm basis on which to choose between them -- although I mentioned in my discussion of Fogelin some reasons I have to be hesitant about strategies like Rorty's and Pears' that involve partitioning the mind. But if we cannot easily choose between these accounts, neither can we endorse all of them, since they are incompatible.

The Puzzles Resolved

I think that the cases of both Pierre and Androvna are cases of in-between believing. It is a mistake to insist on a definite resolution to the question of whether Pierre or Androvna really

have the beliefs that intuition ambivalently attributes and denies to them. And once we let go of the inclination to insist on simple answers to questions about what they believe, the puzzles disappear.

There are actually two steps involved in this approach to the puzzles. The first step is to reject the original Simple Question formulations of the puzzles -- that is, refuse to answer Kripke's insistent question about whether Pierre really does or really does not believe that London is pretty, and, likewise, to refuse to answer the question of whether Androvna really does or really does not believe that she has cancer. So far, the move is not a new one. Both Garrett and Fogelin agree that the question, "Does Pierre believe that London is pretty?" cannot, as it stands, get a simple yes-or-no answer. This point is also argued by Laurence Goldstein (1993) and Graeme Forbes (1994). In the self-deception literature the option of refusing to say that either "yes the self-deceived person believes the unpleasant proposition" or "no she doesn't" is surprisingly uncommon. One sees this view, perhaps, in H. O. Mounce's (1971) paper on the subject, and Mele describes it as an option in a review article on self-deception (1987b), although he neither accepts the idea nor specifically addresses it in his positive work on the topic (1987a).

The more original element of my approach comes with the second step in the resolution of these puzzles. One wants not only to make the negative move just described, but also to develop a *positive* description of the cases at hand. Although

mere recognition of the existence of in-between states of believing may be sufficient to suggest the rejection of the Simple Question in the cases of Pierre and Androvna, a more positive vision of the nature of belief must guide the attempt to give a full satisfactory account of these cases. Here is where my approach diverges from that of Garrett and Fogelin, despite our agreement about the need to reject Kripke's Simple Question.

The difference is that Garrett and Fogelin both allow an all-or-nothing view of belief to re-enter through the back door. Fogelin, although he refuses to say that Pierre, considered as a whole person, either believes or does not believe that London is pretty, does think that Pierre is divisible into *parts* for which simple yes-or-no answers to these questions are appropriate. Similarly Garrett, although he refuses to say that Pierre either believes or does not believe the unqualified proposition that London is pretty, does think that Pierre fully and completely believes the proposition that London, identified in the way associated with the name 'Londres', is pretty, and that Pierre fully and completely believes the proposition that London, identified in the way associated with the name 'London', is not pretty. Both Fogelin and Garrett, then, seem to be seeking some way of carving up affairs so that all legitimate questions about belief *can* get simple yes or no answers. They simply reject the idea that Kripke's original question about Pierre is a legitimate question.

The approach I recommend for describing cases such as Pierre's and Androvna's is the same approach I recommend for describing the multifarious variety of other in-between cases of believing. We should describe the dispositional make-up of the subject at hand, looking both at behavioral dispositions and at phenomenal dispositions; and then we should *stop*. We may, if we wish, note which dispositional patterns match up with which belief stereotypes; we may inquire as to how the subject came to have such a mixed set of dispositions, or how the subject might bring herself better into line with the stereotypes. But these are questions that stand apart from the question of what the subject believes.

There is something approximately right in describing Pierre as believing that London is pretty and in describing Androvna as believing that she has cancer. Both Pierre and Androvna have a number of dispositions that accord with these beliefs, and describing them as having these beliefs can be pragmatically workable to the extent that we can focus our attention and interest on *these* dispositions and explain away with plausible mechanisms other dispositions that accord less well with the stereotypes. At the same time, and for the same reasons, there is something approximately right in describing Pierre as believing that London is not pretty and in describing Androvna as believing that she does not have cancer. But the only completely accurate answer to the question of what Pierre and Androvna believe is an answer that conveys the full mix of their

dispositions without attempting to squeeze them into any of the stereotypes.

Philosophers and psychologists may have felt it necessary to force in-between cases of believing into a simple yes-or-no paradigm because there has been no good picture of belief enabling them to do otherwise. The maneuvers of Kripke, Fogelin, Rorty, and others might then be seen in a Kuhnian (1970) light, as attempts to deal with anomalous data or problem cases by pushing them into the best existing paradigms. My hope is that by presenting a dispositional account of belief and discussing its relation to in-between cases of believing, I have made plausible the claim that there *is* a good alternative to insisting that the only real answers to questions about belief must be of the yes-or-no (or possibly the "degree of belief") variety. Describing a subject as having a divergence of dispositions on a topic is, on my view, *not* settling for less than a full answer to the question of what she believes.

2. What's in a Look?

Major revolutions in a child's cognitive development, like major revolutions in science, do not typically take place all in an instant, but are, as I have repeatedly emphasized, gradual and protracted affairs. If these revolutions can be characterized as changes in (among other things) the child's beliefs, they should be an abundant source of examples of the kind of "in-between" beliefs that are the focus of these chapters. One should positively *expect* periods of in-between believing. I will examine here two cases in which developmental psychologists have been led astray by the inclination to regard the child's knowledge in an all-or-nothing manner. I will begin by exploring Renée Baillargeon's influential views on the infant's understanding of the existence of unperceived objects, and then I will turn to some recent work by Wendy Clements and Josef Perner on the child's understanding of false belief.

The Child's Understanding of Object Permanence

Renée Baillargeon is interested in discovering at what age the child comes to understand that an object observed at two distinct moments in time must also exist in the period between observations. Her work on this topic (e.g., Baillargeon 1987; Baillargeon et al. 1985; Baillargeon and DeVos 1991; Baillargeon, et al. 1990) grows out of a tradition beginning with Piaget (1954). Piaget regards the acquisition of this knowledge about objects as crucial in the development of the concept of "object

permanence", which he sees progressing through several stages between roughly the ages of six and eighteen months. Piaget observed that if a toy in which an infant is interested is removed from view by being placed, in full view of the infant, under a blanket or behind an occluder, children under nine months will not search for it, even though they may have the motor ability to lift blankets and peek behind occluders. It is as though, for the infant, the object no longer existed. Gopnik and Meltzoff (1996), and Harris (1983, 1987) provide interesting reviews of the extensive literature on the development of the object concept.

I will take some time to describe Baillargeon's best-known experiment designed to test the infant's knowledge of object permanence (1987). I will then describe her conclusions from this experiment and provide some arguments against them.

The experimental subjects, 3 1/2- and 4 1/2-month-old infants, were first allowed to handle and were thus familiarized with a 25 x 15 x 5 cm. yellow wooden box with a clown face on it. The infant was then placed before a platform on which a large silver screen lay flat and the yellow box was visible standing upright behind it. The box was then removed and the infant entered the "habituation phase" of the experiment.

In the habituation phase, the large silver screen before the infant was slowly rotated back and forth several times through 180º of arc. The screen began flat on the platform, its top facing the infant, was slowly raised 90º to an upright position, and then was slowly lowered to lay flat against on the platform,

facing away from the infant, having completed 180º of arc. The screen then reversed its path, coming up through 90º and at the end of the cycle lying flat with its top again toward the infant. One cycle took approximately 10 seconds.

The habituation phase acquainted the infant with the motion of the screen and provided a measure against which the infants' looking times at the control and the test events could be measured. A "habituation trial" consisted of a series of cycles, terminating when the infant either (a.) looked away for 2 consecutive seconds after having looked at the display for at least 5 cumulative seconds or (b.) looked at the event for 60 cumulative seconds. Habituation trials were repeated until the infant's looking time on three consecutive trials was 50% or less than her average looking time on the first three trials or until nine cycles were completed, whichever came first.

The infants were then divided into experimental and control conditions. In the experimental condition, the infants were shown two different events, an "impossible event" and a "possible event", in an alternating sequence, until each event had been observed four times. Half the infants saw the impossible event first, and half saw the possible event first.

The "impossible event" began with the screen lying flat toward the infant and the yellow box visible on the platform behind it. The screen was then rotated through 180º of arc, as in the habituation event, while the yellow box was surreptitiously removed so that it would not interfere with the motion of the screen through its last degrees of arc. After

completing the 180°, the screen would reverse its path and the yellow box would be surreptitiously replaced so that at the end of the event the box would be visible again and the screen flat toward the infant. The cycle was then repeated. The end of a trial was determined by the same criteria as the end of a habituation trial. These trials were dubbed the “impossible event” trials because they convey (to an adult) the impression of the screen “impossibly” passing through or squeezing flat the yellow box during its last degrees of arc.

The “possible event” was like the impossible event, except that the screen only rotated through 112° of arc, stopping before hitting the yellow box. The screen then reversed its path to lie flat before the infant with the box visible behind it.

The control conditions were like the experimental conditions, except that the box was absent. Infants in the control condition watched four alternating pairs of 180° and 112° events, just as the infants in the experimental conditions did.

The diagram below, which illustrates some aspects of the conditions just described, is a modified version of a diagram presented in Baillargeon (1987).

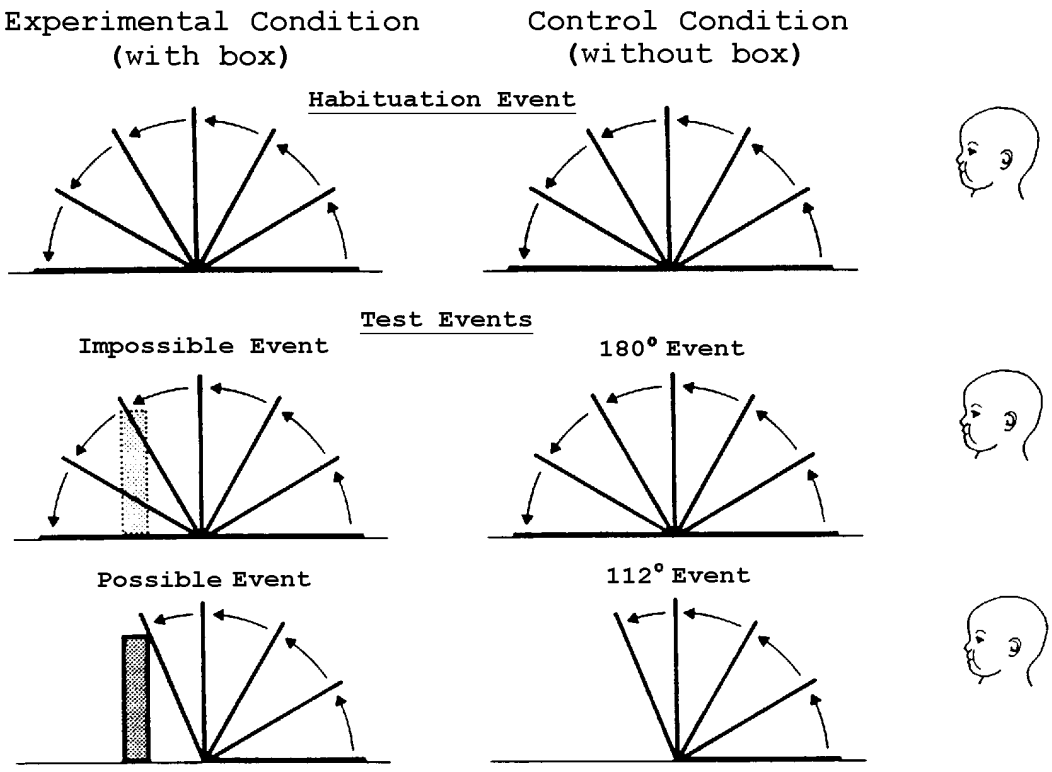

Figure 1. Schematic representation of the habituation and test events shown to the infants in the experimental and control conditions in Baillargeon (1987).

Baillargeon's interest was in the looking times of the infants in the test conditions. This study, as do all "habituation studies", relies on the presupposition that infants will look longer at events that differ more from the event to which they were "habituated" than at those that differ less from the habituation event. Although one could raise methodological questions about this assumption, that is not my plan here. In accordance with the habituation assumption, Baillargeon anticipated that if an infant looked longer at the impossible test events than at the possible ones, that would be because the

infants saw the former as differing more from the habituation event than the latter. This in turn, Baillargeon argues, could only be explained if we assumed that the infant knew that the objects continued to exist even when they were not being perceived. In superficial respects, the 180° impossible event is more like the 180° control event than the 112° possible event is; it is only if one takes into account the apparent "squeezing" or removal of the occluded box that the 180° event seems strange or unique.

Baillargeon found that 4 1/2-month-olds (and "fast habituating" 3 1/2-month-olds) did look significantly longer at the impossible event than at the possible one. This increase in looking time cannot be explained simply by the infants' preferring to watch the screen rotate through 180° over watching it rotate through 112°, because infants in the control condition did not exhibit such a preference. It is natural to suppose, then, that the infants looked longer at the impossible event than the possible one because it violated their expectations about the world.

Baillargeon concludes that, contra Piaget, "infants as young as 3 1/2 months of age already realize that objects continue to exist when occluded" (1987, p. 662). At the same time, she does not deny Piaget's claim that infants' search activities do not reveal such an understanding until the period between nine and eighteen months of age.

The developmental difference between a four-month-old and a nine-month-old is dramatic. The question then arises: If the

infant *really* understands object permanence at four months, why is this understanding not revealed in the child's searching behavior until nine months at least? Baillargeon recognizes this as a difficulty and indicates that the problem may be with the child's means-ends reasoning -- her ability to apply an action to one object (e.g., pull a blanket) to create conditions in which she may apply another action to a different object (e.g., grab the hidden toy). Nevertheless, Baillargeon refers to Piaget's (1952) observations of sequences of behavior in three- and four-month-olds in which an action is applied to one object (e.g., a chain) to produce an effect in another object (e.g., shaking a toy attached to the other end of the chain). Why the latter kind of means-ends reasoning should be available so early and the former kind so late, and what differentiates the two, Baillargeon admits to be "somewhat of a mystery" (1987, p. 663).

With the latter remarks, Baillargeon may be making things harder for herself than she needs to: Piaget doesn't claim really to find means-ends reasoning involving distinct objects until around nine months of age -- the same age at which he discovers search behavior revealing some knowledge of object permanence. The three- or four-month old who pulls a chain to shake an object at the end of it may not see the objects at hand as a system of two separate objects causally related to each other. On the other hand, by six or seven months a child who will not remove an obstacle wholly occluding a desired object will move an obstacle *partly* occluding the desired object (Piaget 1954) and will move a

transparent cover wholly enclosing the desired object (Bower and Wishart 1972), so the problem cannot *simply* be with moving one object to get to another.

In fact, studies on infants in this age range yield starkly divided results on the question of whether infants can reason about objects outside their perceptual fields. On the one hand, when the lights are turned off or the infant rotates her head away from an object, she seems to keep track of its existence (Piaget 1954; Bower and Wishart 1972; Clifton, Rochat, Litovsky, and Perris 1991). Young infants are also able to track the motion of an object as it passes behind an occluder and anticipate the point of reappearance on the other side, sometimes even looking back to the point of disappearance if the object does not reappear in the predicted location (Bower, Broughton, and Moore 1971). There have also been a number of other studies suggesting that infants dishabituate to or look preferentially at events seeming to require that, while occluded, either one object has passed through another or an object has taken a discontinuous path, or in which the number of revealed objects after a period of occlusion is different than an adult would anticipate in the circumstances (Baillargeon 1991; Baillargeon et al. 1985; Spelke et al. 1992; Spelke et al. 1994; Moore et al. 1978; Wynn 1992). On the other hand, a number of studies provide evidence against the idea that infants truly understand that objects continue to exist unperceived. Not only do Piaget's (1954) observations on reaching suggest this, but so also do studies showing that

infants under nine months do not seem disturbed when an object disappears behind one edge of a screen and reappears at the other without making any appearance crossing a gap in the middle of the screen (Moore et al. 1978; but see Baillargeon and Devos 1991); nor do infants show anticipatory reaching for objects on occluded trajectories, though they will reach for objects on visible trajectories (von Hofsten 1994, cited in Gopnik and Meltzoff 1997).

Many of the psychologists conducting such studies, not least Baillargeon, seem committed to arguing one way or another regarding the question of whether the infant genuinely believes that objects continue to exist unperceived. Results pointing in the other direction must then be either discredited or left mysterious.

This area of development would seem to be an excellent example of one on which we ought to say that the infants *neither* truly believe that objects continue to exist unperceived nor truly fail to understand this. Instead, their dispositions on the matter are mixed. Shouldn't one expect infants, in the course of gradual development, to pass through a period like this in any case? It may be time for us to stop beating ourselves over the head looking for a simple yes-or-no answer to the question of whether six-month-olds have an understanding of object permanence. A more useful project would be to determine exactly which of their dispositions point in which directions,

how things came to be this way, and how they change over the course of time.

Implicit Understanding of False Belief?

I conclude with a second case from the developmental literature, Clements' and Perner's paper, "Implicit Understanding of Belief" (1994).

Clements and Perner tested children from 2 1/2 years to 4 1/2 years of age on a variation of the classic false belief experiments performed by Wimmer and Perner (1983) and reported in chapter two. In Clement's and Perner's study, children were introduced to Sam Mouse and his two mouse holes, connected by a V-shaped tunnel. In front of one mouse hole was a red box; in front of the other was a blue box. Children were told the following story, which was simultaneously enacted with cardboard cutouts:

> This is Sam. One day Sam had some cheese for tea. When he looked there was one piece of cheese left but he was too full up to eat it. "I know," he said. "I'll put it in this blue box and I can eat it later." Sam gave a big yawn. "I'm so tired now," he said. He went all the way down the tunnel and went to bed where he fell fast asleep (Clements and Perner 1994, p. 382).

After checking that the child remembered the location of the cheese, the story was continued.

> When Sam had fallen fast asleep, Katie came back from playing outside. As she walked past the blue box, she looked into it and saw the cheese. "Oh look!" she said. "Someone's left a piece of cheese here. I'll put it in the red box and I can eat it later for my tea." So she picked up the cheese and walked, fully visible, across the hill to the other mouse hole where she put the cheese

> in the red box. “I’ll go and see my friend now” she said (Clements and Perner 1994, p. 382).

After asking several questions assuring that the child remembered important facts about the plot, the story was brought to its dramatic conclusion.

> Later on, Sam woke up and gave a big stretch. “I feel very hungry now,” he said. “I’ll go and get the cheese” (Clements and Perner 1994, p. 383).

The experimenter then said, “I wonder where he’s going to look?” and paused for one or two seconds for the child to think about where Sam would look. Throughout this time, the child’s eye movements were recorded on videotape. Finally, the experimenter reminded the child that “Sam wants to get the cheese” and concluded by asking the child two questions: “Which box will he open?” and “Why do you think he will open that box?”

A control group heard much the same story, only with Sam watching while Katie moved the cheese. Half the children heard one of these stories starring Sam Mouse, while half of the children heard a similar story starring Sarah, whose letter was carried by the wind from the upper to the lower balcony of her house.

Children over four tended both to look at the correct box in response to the experimenter’s prompt “I wonder where he’s going to look” and to answer the false belief question correctly. Children under two years, eleven months did exactly the opposite. The interesting results in this study were from the children in the middle age range, from two years eleven months to three years eleven months. Children in this age range typically answered the

false belief question incorrectly but *looked* at the correct box in response to the experimenter's prompt question.

What could explain these results? Clements and Perner reject the hypothesis that in the false belief condition the children are looking at the box in which Sam first left his cheese simply because they are retracing the events of the story, in light of the fact that they don't do this in the control condition and instead look directly at the correct box.[1] Another possibility Clements and Perner reject is that the children's looking reflects tentative hypotheses they momentarily entertain. If this were the case, Clements and Perner argue, the children should have looked at least as frequently at the box they ultimately (and mistakenly) claimed Sam would open as at the other box. Instead, the children look *consistently* at the correct box.

Clements and Perner think the child's eye motions in this experiment reflect some genuine anticipation of Sam's looking in the box in which he originally placed the cheese. Supposing we grant them this, something of a puzzle arises. If the child really understands that Sam will look where he originally left the cheese, why does the child *say* that Sam will look in the other box? Alternately, if the child really *doesn't* understand that Sam will look in the wrong place, how can her eye movements

[1] One possibility Clements and Perner do not consider is that the children retraced the story with their eyes *only* when asked the confusing false belief question. Even such a possibility, however, requires that the children were alert enough to false beliefs that they found the false belief task confusing and the control task simple. Thus, it may still reflect the "implicit" understanding of false belief Clements and Perner argue children this age have.

correctly anticipate the place he will look? The reader may by now scent the likely presence of a mixed dispositional profile.

Clements and Perner could, of course, escape their dilemma by rejecting the Simple Question, but they do not. Instead, they argue that "the eye movements reveal a different type of knowledge" than that revealed by the verbal responses (p. 391). In particular, the eye movements reveal "implicit" knowledge, the verbal responses "explicit" knowledge. Clements and Perner also characterize the difference as one between "nonjudgmental" and "judgmental" knowledge (p. 392). They explain further:

> That is, pure action (i.e. looking in anticipation) is done only on the basis of a representation of reality; that is, one model. But to make a judgment (verbally or gesturally) at least two models are required: One to represent the proposition to be judged (information expressed), and the other to represent the state of the world by which this proposition is to be judged. In other words, to make a judgment is to convey that the verbally or otherwise expressed information (the model of whatever is being proposed) conforms with reality (the other model) (p. 392-393).

Following Karmiloff-Smith (1992), they generalize:

> So, whenever knowledge is acquired in a new domain (acquired procedurally or through abstraction of observed regularities), it becomes first available nonjudgmentally before it can be used to make judgments. For that reason, children in our study are able to anticipate the protagonist's movements correctly with their eyes before they can make a judgment about where he will go (p. 393).

The idea, then, is that major developmental changes in knowledge may be generally first reflected in nonverbal, "nonjudgmental" behavior and only later realized in verbal judgments. And why might this be so? Because verbal judgments are more complicated than "pure action": Pure action requires only that the subject have a correct indicative representation of the world, while

verbal judgment requires combining this with an assessment of the truth or falsity of a proposition expressing these facts about the world.

I am actually sympathetic to the idea that some nonverbal dispositions may be acquired before verbal ones belonging to the same stereotype (although one might imagine this pattern reversed in the case of things taught at school), but Clement's and Perner's view, like many built on all-or-nothing assumptions about belief, adds needless machinery to this observation. What divides "pure actions" not requiring assessment of a proposition from actions like speech that do require such an assessment (and thus two "models") remains something of a mystery. Would opening the correct box instead of naming or pointing to it be a "pure action"? What about interfering with Sam's journey there? The distinction between judgmental and nonjudgmental knowledge must inherit the blurriness of the distinction between pure actions and judgmental ones.

Another problem with Clement's and Perner's view reveals itself as well. Whatever the line between judgmental and nonjudgmental knowledge, conscious verbal assessments must belong to the former category. But even at the same age we see the anticipatory looking, if that's what it is, other signs of judgmental knowledge of false belief are emerging in a limited range of contexts, such as when the child is asked to explain mistaken actions *after* they have occurred (Wellman 1990), and when the child is specifically engaged in the task of "tricking"

someone (Sullivan and Winner 1993). Even, then, if we granted that the distinction between judgmental and nonjudgmental knowledge was a clear one, it would not be motivated by the false belief literature. The picture we see is instead that of a child slowly acquiring the knowledge of false belief: In her early threes, a very few of her dispositions accord with this knowledge, and as she ages, more and more of her dispositions do. It does no good to attempt to salvage all-or-nothing intuitions about belief with the claim that the three-year-old really, fully has one species of knowledge and really, fully lacks another species. The facts are simply not so clean as that.

3. Conclusion

The last three chapters have been occupied with the motivation, explanation, and defense of a novel account of belief, what I have called the *phenomenal dispositional* account of belief. This account arises from the need for an approach to belief that can make sense of *in-between* cases of believing, cases in which the subject is not accurately describable with the everyday "yes-or-no" patterns of belief ascription. The account treats believing as nothing more or less than having dispositions that match stereotypical dispositional profiles. Cases of in-between believing are then treated as cases in which the subject fails to match cleanly with any stereotypical dispositional belief profile.

Several debates in the philosophical and developmental literatures were discussed with the tools provided by the dispositional account, and were shown to profit from the use of those tools. Of particular importance was the ability of the dispositional account to focus its subscriber's interest on problems *other* than trying to extract a simple yes-or-no answer to the question of whether a subject whose dispositional profile is mixed has a particular belief. Trying to force in-between cases of believing into an all-or-nothing mold not only imposes a misleading simplicity on these cases, but also raises a tricky dilemma: On the one hand, if the subject really does fully and completely have the belief, how it is possible that she does not manifest it in a wide variety of circumstances? On the other

hand, if the subject does not really have the belief, how can it be that she seems sometimes to act on the basis of the knowledge denied her? It is tempting to try to escape this dilemma by inventing mental machinery, as we have seen in the cases of Fogelin, Rorty, and Clements and Perner. The danger in this move is not that dividing the mind or introducing different faculties of believing is in itself a mistake, but rather that its postulation in these cases is only as justified as the resolution to describe these cases in an all-or-nothing manner.

Besides the danger of insisting too adamantly on discovering simple yes-or-no answers to questions about what a subject believes, however, is the converse danger -- that of giving up too quickly in finding such answers. In chapter five I outlined a primary reason for seeking such yes-or-no answers: People generally conform fairly well to the stereotypes, and evidence pointing toward a mixed dispositional profile will often sort itself out clearly in favor of one stereotype or another. It is important to distinguish cases in which a person only *seems* to have mixed dispositions from cases in which the nonconformity is genuine. Good judgment will have to be our guide in deciding when to concede the presence of a genuinely mixed dispositional profile and thus to give up on finding simple yes-or-no answers to what the subject believes. The judgment is complicated by the presence of more than simply epistemic factors. The yes-or-no approach also has the advantage of simplicity, which may in some contexts outweigh the increased accuracy of more detailed

dispositional descriptions when the subject fairly closely matches one stereotype or another; and furthermore, insistence on simple yes-or-no questions about belief may also serve the purpose of motivating both ourselves and others to conform to societially necessary dispositional stereotypes, as suggested in chapter six.

Besides having these reasons for insisting on yes-or-no answers to questions about belief, philosophers and psychologists may have felt it necessary to force in-between cases of believing into a simple yes-or-no paradigm because there has been no good scientific alternative allowing one to do otherwise. The maneuvers of Kripke, Fogelin, Baillargeon, and the others might then be seen generously, in a Kuhnian (1970) light, as attempts to deal with anomalous data or problem cases by pushing them into the best existing paradigm. One could hardly expect a good scientist to do otherwise. My hope is that these chapters have convinced the reader that there *is* a good scientific alternative to insisting that the only "real" answers to questions about belief must be of the yes-or-no (or possibly the "degree of belief") variety -- and that describing a subject as having a divergence of dispositions on a topic is not settling for less and provides no hindrance to scientific research. It is worth noting in this regard that neural net models of cognition (classically described in Rumelhart et al. 1986 and McClelland et al. 1986) seem to allow quite naturally a broad range of in-between responses and dispositional mixes, and that if neural net

models find broad use in understanding human cognition, an account of cognition that can handle these “in-betweenish” features of neural nets will be necessary.

So much for the pragmatic benefits of the account. The ontological dimension of the account is, I hope, conservative and widely acceptable. There is, of course, some talk about properties, dispositions, and stereotypes, but I do not believe it has been necessary to take any controversial stands on these matters. I have claimed that in-between cases of believing are common, and I have provided a number of examples of such cases. While any individual example may itself be controversial as an instance of in-between believing, what is important to my position in not any individual case, but rather the overall impression I sought to create of the ubiquity of such in-between cases.

One ontological claim, however, is crucial to my account and at the same time potentially controversial. It is the claim that once one has fully described a subject’s dispositional profile and compared that profile to the relevant stereotypes, one has exhausted everything we can know about what that subject believes on the topic. There is no *further* fact of the matter, apart from facts about the subject’s dispositional profile, about what the subject “really” believes -- or at least no fact we can presently discover. It is unclear what would count as a discovery of such a fact (unless we consider the possibility of science eventually developing in the direction suggested at the end of chapter six),

and accounts like the one offered here show, I think, that we can run philosophy and the sciences without appeal to such facts. Occam's razor, then, recommends leaving them out of our ontology.

One can view the project of this chapter as a revamping of the old Rylean dispositionalist view of belief, with a new emphasis on the phenomenal aspects of the account. This project is quite timely in its way. The 1990's have seen a resurgence of academic interest in the phenomenal aspects of mind (Searle 1992 is an excellent example), and I should not be surprised to see quite a number of mid-century views reincarnated with a phenomenal twist.

Chapter Eight

Conclusion

In this dissertation, I have woven together philosophical issues with issues in empirical developmental psychology, in hopes of producing a work that may usefully be read by people in both disciplines. My primary goal has been the clarification of three concepts employed centrally in the two disciplines, the concepts of *theory, representation,* and *belief*. I have treated these concepts, and the words with which we label them, as practical tools that philosophers and psychologists use in understanding the human (or animal) mind. As tools of this sort, I have argued that they should be evaluated functionally, in terms of their ability to assist us in reaching an informed understanding of the mind, and that we should feel free to modify them in whatever way best helps us achieve this goal. Adopting such an approach, I have proposed novel accounts of the concepts of *theory* and *belief*, and I have shown some of the dangers of an inconsistent approach to the concept of *representation*.

In my approach to the concept of a theory, I had two practical applications in mind. Primarily, I wanted to develop an account of theories that would be useful in clarifying the developmental debate over the extent to which the cognitive development of children should be described as “theoretical.” Secondarily, I wanted to develop an account of theories that

applied equally to the informal theories of everyday life and the technical theories of advanced science, on the assumption that there is some important continuity between the two types of "theory" that might be revealed by such an account. If the first goal were to be met, it seemed the second would also have to be met, since if it makes any sense at all to debate the extent to which children are theorizers, the debate must depend on an understanding of theories that includes the informal theories of everyday life. The resulting account connected theories tightly with the satifaction of a "drive to explain": Theories were necessarily to be evaluated in terms of their capacity to generate good explanations on the topic at hand, and a person was said to subscribe to a theory when she was disposed to employ it in explanations, or at least for the resolution of "explanation-seeking curiosity." If such an account of theories is acceptable for the purposes of the debate over the "theory theory" in developmental psychology, then, I argued, we ought to see patterns of affect and arousal indicative of the emergence and resolution of explanation-seeking curiosity in the kinds of puzzling situations that would, according to the theory theory, stimulate development by forcing the generation of new theories. Thus, I suggested, affect and arousal offer a new domain of evidence against which the theory theory should be tested.

My goals in discussing the concept of representation were also multiple. One of those goals can be thought of as primarily developmental and another as primarily philosophical. The philosophical goal was a clarification of the difference between

two types of account of representation -- one I labeled ‘contentive,’ the other ‘indicative’ -- a difference that, I argued, has not always been clearly noticed, even by philosophers instrumental in the development of these accounts (such as Stampe and Fodor). The developmental goal was the diagnosis of the failure of a certain research program in developmental psychology, the existence of which, I argued, depended on assumptions that only seemed to be justified given a conflation of these two types of representation. In particular, I argued that the research program in question depended on the assumption that the child’s understanding of desire must undergo a tranformation at age four analogous to the child’s transformation in understanding belief at that age. In lieu of the vain search for such a transformation, I suggested another direction for research on the child’s understanding of representation, involving the child’s understanding of representational art. A third, overaching goal also motivated my discussion of representation. As is suggested by the title of the chapter on representation, I see the chapter as a case study of how philosophical errors can be harmful to empirical research. Perhaps if enough such cases are elaborated, that will help motivate people in empirical fields to seek out philosophical understanding in developing their more theoretically-loaded experiments and views. Also, it may help strengthen the conviction of some philosophers that there is interesting philosophical work to be done in the interpretation and motivation of empirical research.

My discussion of the concept of belief covers four chapters of the dissertation, and is the most variously motivated. Chapter two was primarily motivated by a concern over what seems to be a common form of philosophical myopia: the tendency of some philosophers to dictate to academics in other fields the use of certain words and concepts without sufficient concern for the interests of researchers in those fields in using those concepts. In particular, I argued that developmental psychology and cognitive ethology would be damaged by insistence on avoiding the ascription of beliefs to infants and non-human animals without language. Especially given the failure of arguments attempting to establish the gross inapplicability of that concept to such creatures, I argued that we ought to consider it a condition of acceptability of a general-purpose account of belief that it apply to infants and at least some non-human animals.

In the fifth, sixth, and seventh chapters of the dissertation, I offered a novel analysis of the concept of belief. I suggested that we think of believing that P as matching, to an appropriate degree and in appropriate respects, the “dispositional stereotype” for believing that P. Since the term ‘belief’ is already common coin in both philosophy and psychology, it is useful to develop an account of belief that matches fairly well in extension with existing usage: Most of what philosophers and psychologists consider to be cases of believing should turn out to be cases of believing, under the new definition, and most of what they consider not to be cases of believing should turn out not to be. Otherwise, integration of

the account into existing theoretical structures might cause unnecessary difficulties. The account I offered satisfies this practical condition. In addition, the account has, I believe, the pragmatic virtues of clarity and simplicity. However, the primary virtue that I claimed for the account over and above other accounts was its facility in handling "in-between" cases of believing, cases in which the subject in not accurately described either as completely believing that something is the case or as completely failing to believe it. Although some such in-between cases can be described well enough with Bayesian degrees of belief, I reviewed a wide variety of cases for which this was not so and upon which typical philosophical and psychological approaches to belief have foundered. In chapter seven, I explored four such cases in depth, and I showed how a dispositional account of belief allows us fruitfully to describe such cases and move on with our philosophical and psychological work.

Conceptual analysis is one of the most fundamental tasks of philosophy. Yet, since concepts are ours for the remaking, there is always an indefinite variety of possible analyses of any particular concept. Without particular practical goals in mind against which to measure the success of our analyses, philosophical debates can seem to be ungrounded and empty. Connecting philosophical work with the empirical sciences not only gives it a relevance beyond the sometimes insular world of the philosophical journals, but also can provide the very ground that makes philosophical inquiry meaningful.

Works Cited

Anscombe, G. E. M. (1957). *Intention.* Oxford: Basil Blackwell.

Armstrong, David M., (1968). *A materialist theory of mind.* London: Routledge and Kegan Paul.

Arnheim, Rudolf (1974). *Art and visual perception.* Berkeley: University of California.

Astington, Janet W. (1991). Intention in the child's theory of mind. In Douglas Frye and Chris Moore (eds.), *Children's theories of mind.* Hillsdale, NJ: Erlbaum.

Astington, Janet W. (1993). *The child's discovery of mind.* Cambridge, Mass.: Harvard.

Astington, Janet W. and A. Gopnik (1988). Knowing you've changed your mind: Children's understanding of representational change. In J. Astington, P. Harris, and D. Olson (eds.), *Developing theories of mind.* Cambridge: Cambridge.

Astington, Janet W. and A. Gopnik (1991). Developing understanding of desire and intention. In A. Whiten (ed.), *Natural theories of mind: Evolution, development, and simulation of everyday mindreading.* Oxford: Basil Blackwell.

Audi, Robert (1982). Self-deception, action, and will. *Erkenntnis, 18,* 133-158.

Audi, Robert (1985). Self-deception and rationality. In M. W. Martin (ed.), *Self-deception and self-understanding.* Lawrence, KA: University Press of Kansas.

Audi, Robert (1994). Dispositional beliefs and dispositions to believe. *Nous, 28,* 419-434.

Austin, John L. (1956). A plea for excuses. *Proceedings of the Aristotelian Society, 57*.

Bach, Kent (1981). An analysis of self-deception. *Philosophy and Phenomenological Research, 46,* 351-370.

Baillargeon, Renée (1987). Object permanence in 3 1/2- and 4 1/2-month-old infants. *Developmental Psychology, 23,* 655-664.

Baillargeon, Renée and J. DeVos (1991). Object permanence in young infants: Further evidence. *Child Development, 62,* 1227-1246.

Baillargeon, Renée, M. Graber, J. DeVos, and J. Black (1990). Why do young infants fail to search for hidden objects? *Cognition, 26,* 255-284.

Baillargeon, Renée, E. S. Spelke, and S. Wasserman (1985). Object permanence in five-month-old infants. *Cognition, 20,* 191-208.

Bartsch, Karen and H. M. Wellman (1989). Young children's attribution of action to beliefs and desires. *Child Development,* 60, 946-964.

Bartsch, Karen and H. M. Wellman (1995). *Children talk about the mind.* New York: Oxford.

Bates, Elizabeth and J. Elman (1993). Connectionism and the study of change. In M. J. Johnson (ed.), *Brain development and cognition: A reader.* Oxford: Blackwell.

Bates, Elizabeth, J. Elman, and A. Karmiloff-Smith (1995). Why connectionism? Forthcoming.

Beatty, John (1981). What's wrong with the received view of evolutionary theory? In P. D. Asquith and R. N. Giere (eds.) *PSA 1980 Vol.2.* East Lansing, MI: Philosophy of Science Association, 397-426.

Beatty, John (1987). On behalf of the semantic view. *Biology and Philosophy, 2,* 17-23.

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

Bishop, John (1980). More thought on thought and talk. *Mind, 89,* 1-16.

Black, Max (1962). *Models and metaphors.* Ithaca, NY: Cornell University Press.

Block, Ned (1978). Troubles with functionalism. In C. Wade Savage (ed.), *Minnesota Studies in Philosophy of Science, 9.* Minneapolis: University of Minnesota Press, 261-325.

Bower, T. G. R., J. Broughton, and M. K. Moore (1971). The development of the object concept as manifested in tracking behavior of infants between 17 and 20 weeks of age. *Journal of Experimental Child Psychology, 11,* 182-193.

Bower, T. G. R., andJ. G. Wishart (1972). The effects of motor skill on object permanence. *Cognition, 1,* 165-172.

Bretherton, Inge and M. Beeghly (1982). Talking about internal states: The acquisition of an explicit theory of mind. *Developmental Psychology, 18,* 906-921.

Bromberger, Sylvain (1962). An approach to explanation. In R. S. Butler (ed.), *Analytical philosophy, Second series.* Oxford: Basil Blackwell, p. 72-105.

Bruner, Jerome (1992). The narrative construction of reality. In H. Beilin and P. Pufall (eds.), *Piaget's theory: Prospects and possibilities.* Hillsdale, N.J.: Lawrence Erlbaum.

Burge, Tyler (1979). Individualism and the mental. P. French T. E. Uehling, Jr., and H. Wettstein (eds.), *Midwest Studies in Philosophy, 4: Studies in Metaphysics.* Minneapolis: University of Minnesota Press.

Carey, Susan (1985). *Conceptual change in childhood.* Cambridge, MA: MIT Press.

Carnap, Rudolf (1962). *Logical foundations of probability, 2d ed.* Chicago: University of Chicago.

Carnap, Rudolf (1936/1954). *Testability and meaning*. New Haven: Whitlock's.

Carnap, Rudolf (1966). *Philosophical foundations of physics.* Ed. Martin Gardner. New York: Basic Books.

Cartwright, Nancy (1983). *How the laws of physics lie.* New York: Oxford UP.

Case, Robbie (1985). *Intellectual development: Birth to adulthood.* Orlando, FL: Academic Press.

Case, Robbie and Y. Okamoto (1996). The role of central conceptual structures in the development of children's thought, *Monographs of the Society for Research in Child Development, 61.*

Charman, Tony and S. Baron-Cohen (1993). Drawing development in autism: The intellectual to visual-realism shift. *British Journal of Developmental Psychology, 11,* 171-185.

Charman, Tony and S. Baron-Cohen (1992). Understanding drawings and beliefs: A further test of the metarepresentation theory of autism: A research note. *Journal of Child Psychology and Psychiatry, 33,* 1105-1112.

Chisholm, Roderick M. (1957). *Perceiving: A philosophical study.* Ithaca, NY: Cornell.

Chomsky, Noam (1975). *Reflections on language.* New York: Random House.

Chomsky, Noam (1980). *Rules and representations*. New York: Columbia.

Churchland, Paul M. (1981). Eliminative materialism and the propositional attitudes. *The Journal of Philosophy, 78,* 67-90.

Churchland, Paul M. (1990). Cognitive activity in artificial neural networks. In D. N. Osherson and E. E. Smith (eds.), *Thinking: An invitation to cognitive science.* Cambridge, Mass.: MIT.

Clark, Andrew (1993). *Associative engines: Connectionism, concepts, and representational change.* Cambridge, Mass.: MIT.

Clements, Wendy A. and J. Perner (1994). Implicit understanding of belief. *Cognitive Development, 9,* 377-395.

Clifton, Rachel K., P. Rochat, R. Litovsky, and E. Perris (1991). Object representation guides infants' reaching in the dark. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 323-329.

Cooper, Joel and R. H. Fazio (1984). A new look at dissonance theory. *Advances in Experimental Social Psychology, 17,* 229-266.

Darwin, Charles (1911). *The formation of vegetable mould through the action of worms, with observations on their habits*. New York: D. Appleton.

Davidson, Donald (1973/1984). Radical interpretation. *Dialectica, 27,* 313-328. Reprinted in D. Davidson, *Inquiries into truth and interpretation*. Oxford: Clarendon.

Davidson, Donald (1974/1984). Belief and the basis of meaning. *Synthese, 27,* 309-323. Reprinted in D. Davidson, *Inquiries into truth and interpretation*. Oxford: Clarendon.

Davidson, Donald (1975/1984). Thought and talk. In S. Guttenplan (ed.), *Mind and language*. Oxford: Oxford UP, 1975. Reprinted in D. Davidson, *Inquiries into truth and interpretation*. Oxford: Clarendon.

Davidson, Donald (1982a). Paradoxes of irrationality. In Richard Wollheim and J. Hopkins (eds.), *Philosophical essays on Freud.* Cambridge: Cambridge University Press.

Davidson, Donald (1982b). Rational animals. *Dialectica, 36*, 317-327.

Davidson, Donald (1984). Inquiries into truth and interpretation. Oxford: Clarendon.

Davidson, Donald (1985a). Communication and convention. In M. Dascal (ed.), *Dialogue,* 11-25.

Davidson, Donald (1985b). Deception and division. In Ernest LePore and B. McLaughlin (eds.), *Actions and events.* New York: Basil Blackwell.

Davidson, Donald (1986). A nice derangement of epitaphs. In E. Lepore (ed.), *Truth and interpretation: Perspectives on the philosophy of Donald Davidson*. Cambridge: Blackwell.

Davidson, Donald (1987). Knowing one’s own mind. *Proceedings and Addresses of the American Philosophical Association, 60.*

Davidson, Donald (1994). The social aspect of language. In B. McGuinness (ed.), *The philosophy of Michael Dummett*.

de Waal, Frans (1989). *Peacemaking among primates.* Cambridge, Mass.: Harvard.

DeLoache, Judy S. (1989a). The development of representation in young children. *Advances in Child Development and Behavior, 22,* 1-39.

DeLoache, Judy S. (1989b). Young children's understanding of the correspondence between a scale model and a larger space. *Cognitive Development, 4,* 121-139.

DeLoache, Judy S. (1991). Symbolic functioning in very young children: Understanding of pictures and models. *Child Development, 62,* 736-752.

DeLoache, Judy S. (1995). Early symbol understanding and use. The Psychology of Learning and Motivation, 33, 65-114

DeLoache, Judy S. and C. M. Smith (forthcoming). Early symbolic representation. In I. Sigel (ed.), *Theoretical perspectives in the concept of representation*. Hillsdale, N.J.: Erlbaum.

Demos, Raphael (1960). Lying to oneself. *Journal of Philosophy, 57,* 588-595.

Dennett, Daniel C. (1978a). Beliefs about beliefs. *Behavioral and Brain Sciences, 1,* 568-570.

Dennett, Daniel C. (1978b). *Brainstorms*. Cambridge, MA: MIT Press.

Dennett, Daniel C. (1983). Intentional systems in cognitive ethology: The "Panglossian paradigm" defended. *Behavioral and Brain Sciences, 6,* 343-390.

Dennett, Daniel C. (1987). *The intentional stance.* Cambridge, Mass.: MIT Press.

Dennett, Daniel C. (1991a). *Consciousness explained.* Boston: Little, Brown.

Dennett, Daniel C. (1991b). Real patterns.

Descartes, René (1637/1980). *Discourse on Method*. In *Discourse on Method and Meditations on First Philosophy*. Trans. D. A. Cress. Indianapolis: Hackett.

di Sessa, Andrea A. (1988). Knowledge in pieces. In G. Forman and P. B. Pufall (eds.), *Constructivism in the computer age*. Hillsdale, N.J.: Lawrence Erlbaum.

Doherty, Martin and J. Perner (1997). Metalinguistic awareness and theory of mind: Just two words for the same thing? Unpublished manuscript.

Downes, Stephen M. (1993). The importance of models in theorizing: A deflationary semantic view. *PSA 1992 Vol. 1,* East Lansing, MI: Philosophy of Science Association, 142-153.

Dretske, Fred (1988). *Explaining behavior.* Cambridge, Mass.: MIT Press.

Dretske, Fred (1993). The nature of thought. *Philosophical Studies, 70,* 185-199.

Dretske, Fred (1995). *Naturalizing the mind.* Cambridge, Mass.: MIT Press.

Dupre, John (1993). The disorder of things. Cambridge, Mass.: Harvard.

Feyerabend, Paul (1975/1988). *Against method, rev. ed.* London: Verso.

Field, H. H. (1978). Mental representation. *Erkenntnis, 13,* 9-61.

Fine, Arthur (1989). Do correlations need to be explained? In J. T. Cushing and E. McMullin (eds.), *Philosophical consequences of quantum theory: Reflections on Bell's theorem.* Notre Dame, Ind.: University of Notre Dame.

Fischer and Bidell (1991). Constraining nativist inferences about cognitive capacities. In S. Carey and S. Gelman (eds.), *Epigenesis of mind: Essays on Biology and Cognition.* Hillsdale, N.J.: Erlbaum.

Flavell, John H. (1988). The development of children's knowledge about the mind: From cognitive connections to mental representations. In J. W. Astington, P. L. Harris, and D. R. Olson (Eds.), *Developing theories of mind,* 244-267. New York: Cambridge University Press.

Flavell, John. H., E. R. Flavell, and F. L. Green (1983). Development of the appearance-reality distinction. *Cognitive Psychology, 15,* 95-120.

Flavell, John. H., E. R. Flavell, and F. L. Green (1989). Transitional period in the development of the appearance-reality distinction. *International Journal of Behavioral Development, 12,* 509-526.

Flavell, John H., E. R. Flavell, F. L. Green, and L. J. Moses (1990). Young children's understanding of fact beliefs versus value beliefs. *Child Development, 61,* 915-928.

Flavell, John H., F. L. Green, and E. R. Flavell (1986). Development of knowledge about the appearance-reality distinction. *Monographs of the Society for Research in Child Development #212, 51.*

Flavell, John H., F. L. Green, K. E. Wahl, and E. R. Flavell (1987). The effects of question clarification and memory aids on young children's performance on appearance-reality tasks. *Cognitive Development, 2,* 127-144.

Fodor, Jerry A. (1968). *Psychological explanation.* New York: Random House.

Fodor, Jerry A. (1975). *The language of thought.* Cambridge, Mass.: Harvard.

Fodor, Jerry A. (1981). *Representation: Philosophical essays on the foundations of cognitive science.* Cambridge, Mass.: MIT Press.

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

Fodor, Jerry A. (1984). Semantics, Wiconsin style. *Synthese, 59,* 231-250.

Fodor, Jerry A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind.* Cambridge, Mass.: MIT.

Fodor, Jerry A. (1990). *A theory of content.* Cambridge, Mass.: MIT.

Fodor, Jerry A. (1991). Fodor's guide to mental representation: The intelligent auntie's vade-mecum. In J. Greenwood (ed.),

The future of folk psychology: Intentionality and cognitive science. Cambridge: Cambridge.

Fodor, Jerry A. (1992). A theory of the child's theory of mind. *Cognition, 44,* 283-296.

Fodor, Jerry A. (1994). *The elm and the expert: Mentalese and its semantics.* Cambridge, MA: MIT.

Fogelin, Robert J. (1994). Pierre, Saul, Ruth, and Bob and a puzzle about belief. In W. Sinnott-Armstrong (ed.), *Modality, Morality, and Belief.* New York: Cambridge.

Forbes, Graeme (1994). Donnellan on a puzzle about belief. *Philosophical Studies, 73,* 169-180.

Forguson, Lynd and A. Gopnik, A. (1988). The ontogeny of common sense. In J. Astington, P. Harris, and D. Olson (eds.), *Developing theories of mind.* Cambridge: Cambridge.

Freeman, Norman H. (1980). *Strategies of representation in young children.* London: Academic.

Freeman, Norman H. and H. Lacohée (1995). Making explicit 3-year-olds' implicit competence with their own false beliefs. *Cognition, 56,* 31-60.

Freeman, Norman H., C. Lewis, and M. J. Doherty (1991). Preschoolers' grasp of a desire for knowledge in false-belief prediction: Practical intelligence and verbal report. *British Journal of Developmental Psychology, 9,* 139-157.

Freud, Sigmund (1977). *Five lectures on psychoanalysis.* Trans. J. Strachey. New York: W.W. Norton.

Friedman, Michael (1974/1988). Explanation and scientific understanding. In Joseph C. Pitt (ed.), *Theories of explanation.* Oxford: Oxford UP.

Friend, Margaret and T. L. Davis (1993). Appearance-reality distinction: Children's understanding of the physical and affective domains. *Developmental Psychology, 29,* 907-914.

Frye, Douglas and C. Moore, eds. (1991). *Children's theories of mind.* Hillsdale, N.J.: Erlbaum.

Gardner, Howard and D. Wolf (1987). Symbolic products of early childhood. In D. Görlitz and J. F. Wohlwill (eds.),

Curiosity, imagination, and play: On the development of spontaneous cognitive and motivational processes. Hillsdale, N.J.: Lawrence Erlbaum.

Garrett, Richard (1991). Putnam on Kripke's puzzle. *Erkenntnis, 34,* 271-286.

Gelman, Susan A. and J. D. Coley (1991). Language and categorization: The acquisition of natural kind terms. In S. A. Gelman and J. P. Byrnes (eds.), *Perspectives on language and thought: Interrelations in development.* New York: Cambridge.

Giere, Ronald (1988). *Explaining science: A cognitive approach.* Chicago: University of Chicago.

Giere, Ronald (1989). *Understanding scientific reasoning, 3d ed.* Fort Worth: Harcourt Brace.

Goldstein, Lawrence (1993). The fallacy of the simple question. *Analysis, 53,* 178-181.

Golomb, Claire (1992). *The child's creation of a pictorial world.* Berkeley: University of California.

Gomez, Juan C. (1994). Mutual awareness in primate communication: A Gricean approach. In S. T. Parker, R. W. Mitchell, and M. L. Boccia (eds.), *Self-awareness in animals and humans.* Cambridge: Cambridge.

Gopnik, Alison (1988). Conceptual and semantic development as theory change: The case of object permanence. *Mind and Language, 3,* 197-216.

Gopnik, Alison (1990). Developing the idea of intentionality: Children's theories of mind. *Canadian Journal of Philosophy, 20,* 89-114.

Gopnik, Alison and J. W. Astington (1988). Children's understanding of representational change and its relation to the understanding of the false belief and the appearance-reality distinction. *Child Development, 59,* 26-37.

Gopnik, Alison and A. Meltzoff (1993). Words and thoughts in infancy: The specificity hypothesis and the development of

categorization and naming. In *Advances in Infancy Research, 8,* 217-249.

Gopnik, Alison and A. Meltzoff (1997). *Words, thoughts, and theories.* Cambridge, Mass.: MIT.

Gopnik, Alison and E. Schwitzgebel (1996). Whose concepts are they, anyway? Forthcoming.

Gopnik, Alison and V. Slaughter (1991). Young children's understanding of changes in their mental states. *Child Development, 62,* 98-110.

Gopnik, Alison and H. Wellman (1992). Why the child's theory of mind really is a theory. *Mind and Language, 7,* 145-171.

Griesemer, James R. (1990). Modeling in the museum: On the role of remnant models in the work of Joseph Grinnell. *Biology and Philosophy, 5,* 3-36.

Hadwin, Julie and J. Perner (1991). Pleased and surprised: Children's cognitive theory of emotion. *British Journal of Developmental Psychology, 9,* 215-234.

Hala, Suzanne, M. Chandler, and A.S. Fritz (1991). Fledgling theories of mind: Deception as a marker of three-year-olds' understanding of false belief. *Child Development, 62,* p. 83-97.

Hampshire, Stuart (1950). The concept of mind. *Mind, 59,* 237-255.

Harman, Gilbert (1978). Studying the Chimpanzee's Theory of Mind. *Behavioral and Brain Sciences, 1*.

Harman, Gilbert (1986). *Change in view: Principles of reasoning.* Cambridge, Mass.: MIT.

Harris, Paul (1983). *Handbook of child psychology, vol. 2*.

Harris, Paul (1987). *Handbook of infant perception, vol. 2*.

Harris, Paul L., C. N. Johnson, D. Hutton, G. Andrews, and T. Cooke (1989). Young children's theory of mind and emotion. *Cognition and Emotion, 3,* 379-400.

Hearne, Vicki (1982). *Adam's task.* New York: Random House.

Heil, John (1992). *The nature of true minds*. Cambridge: Cambridge.

Hempel, Carl G. (1952). Fundamentals of concept formation in empirical science. From the *International Encyclopedia of Unified Science, Vol. 2,* Number 7, Ed. Otto Neurath. Chicago: Universityof Chicago.

Hempel, Carl G. (1965). *Aspects of scientific explanation and other essays in the philosophy of science*. New York: Free Press.

Hempel, Carl G. and P. Oppenheim (1948). Studies in the logic of explanation. *Philosophy of Science, 15,* 567-579.

Hewson, Peter A. and M. A. Hewson(1984). The role of conceptual conflict in conceptual change and the design of instruction. *Instructional Science, 13,* 1-13.

Hogrefe, G.-Jürgen, H. Wimmer, and J. Perner (1986). Ignorance versus false belief: A developmental lag in attribution of epistemic states. *Child Development, 57,* 567-582.

Hughes, R. I. G. (1989). Bell's theorem, ideology, and structural explanation. In J. T. Cushing and E. McMullin, eds., *Philosophical consequences of quantum theory: Reflections on Bell's theorem.* Notre Dame, Ind.: University of Notre Dame.

Humberstone, I. L. (1992). Direction of Fit. *Mind, 101,* 59-83.

Humphreys, Paul (1989). Scientific explanation: The causes, some of the causes, and nothing but the causes. In P. Kitcher and W.C. Salmon, eds., *Minnesota studies in the philosophy of science, vol. XIII: Scientific explanation.* Minneapolis: Universityof Minnesota Press.

Jeffrey, Richard C. (1983). *The logic of decision.* Chicago: University of Chicago.

Jeffrey, Richard C. (1992). *Probability and the art of judgment.* Cambridge: Cambridge University Press.

Karmiloff-Smith, Annette (1984). Children's problem solving. In M.E. Lamb, A.L. Brown, and B. Rogoff, eds., *Advances in Developmental Psychology, Vol. 3.* New Jersey: Erlbaum, 39-90.

Karmiloff-Smith, Annette (1988). The child is a theoretician, not an inductivist. *Mind and Language, 3,* p. 183-195.

Karmiloff-Smith, Annette (1990). Constraints on representational change: Evidence from children's drawing. *Cognition*, *34,* 57-83.

Karmiloff-Smith, Annette (1992). *Beyond modularity: A developmental perspective on cognitive science.* Cambridge, Mass.: MIT.

Keil, Frank (1991) Theories, concepts and the acquisition of word meaning. In S. A. Gelman and J. P. Byrnes (eds.), *Perspectives on language and thought: Interrelations in development.* New York: Cambridge.

Keller, Evelyn F. and E. A. Lloyd, eds. (1992). *Keywords in evolutionary biology.* Cambridge, Mass.: Harvard.

Kitcher, Philip (1988). The child as parent of the scientist. *Mind and Language, 3,* p. 217-228.

Kitcher, Philip (1989). Explanatory unification and the causal structure of the world. In P. Kitcher and W. C. Salmon (eds.), *Minnesota studies in the philosophy of science, vol. XIII: Scientific explanation.* Minneapolis: University of Minnesota.

Köhler, Wolfgang (1926). *The mentality of apes, 2d ed*. Trans. E. Winter. New York: Harcourt Brace.

Kripke, Saul A. (1979). A puzzle about belief. In A. Margalit (ed.), *Meaning and Use.* Dordrecht: Reidel, 239-283.

Kuhn, Thomas (1962/1970). *The structure of scientific revolutions.* Second Ed. Chicago: University of Chicago.

Kuhn, Thomas (1977). Objectivity, value judgment, and theory choice. In *The essential tension*. Chicago: University of Chicago.

Lakoff, George and M. Johnson (1980). *Metaphors we live by.* Chicago: University of Chicago.

Lecours, André R. and Y. Joanette (1980). Linguistic and other psychological aspects of paroxysmal aphasia. *Brain and Language, 10,* 1-23.

Leslie, Alan M. (1987). Pretense and representation: The origins of "theory of mind." *Psychological Review, 94,* 416-426.

Leslie, Alan M. (1988). Some implications of pretense for mechanisms underlying the child's theory of mind. In J. W. Astington, P. L. Harris, and D. R. Olson (eds.), *Developing theories of mind*. New York: Cambridge UP.

Leslie, Alan M. (1994a). Pretending and believing: Issues in the theory of ToMM. *Cognition, 50,* 211-238.

Leslie, Alan M. (1994b). ToMM, ToBy, and agency: Core architecture and domain specificity. In L. A. Hirschfeld and S. A. Gelman, eds., *Mapping the mind: Domain specificity in cognition and culture.* Cambridge: Cambridge UP.

Lewis, Charlie and A. Osborne (1990). Three-year-olds' problems with false belief: Conceptual deficit or linguistic artifact? *Child Development, 61,* 1514-1519.

Lewis, David (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy, 50,* 249-258.

Lewis, David (1980). Mad pain and Martian pain. In Ned Block (ed.), *Readings in the philosophy of psychology, vol. 1.* Cambridge, Mass: Harvard.

Lewis, David (1986a). Causal explanation. In *Philosophical papers, vol. 2.* Oxford: Oxford

Lewis, David (1986b). Causation. In *Philosophical papers, vol. 2.* Oxford: Oxford

Liben, Lynn S. and R. M. Downs (1989). Understanding maps as symbols: The development of map concepts in children. *Advances in Child Development and Behavior, 22,* 145-199.

Lillard, Angeline S. and J. H. Flavell (1992). Young children's understanding of different mental states. *Developmental Psychology, 28,* 626-634.

Lloyd, Elisabeth A. (1987). Response to Sloep and van der Steen. *Biology and Philosophy, 2,* 23-26.

Lloyd, Elisabeth A. (1988). *The structure and confirmation of evolutionary theory.* Princeton, N.J.: Princeton.

Lloyd, Elisabeth A. (1995). Objectivity and the double standard for feminist epistemologies. *Synthese, 104,* 351-381.

Locke, John (1690/1975). *An essay concerning human understanding.* Ed. P. H. Nidditch. Oxford: Oxford.

Mackie, John L. (1973). *Truth, probability, and paradox.* Oxford: Oxford.

Mackie, John L. (1974). *The cement of the universe: A study of causation.* Oxford: Oxford.

Malcolm, Norman (1942). Moore and ordinary language. In P. A. Schilpp, ed., *The philosophy of G. E. Moore.* Evanston, Ill.: Northwestern.

Malcolm, Norman (1973). Thoughtless brutes. *Proceedings and addresses of the American Philosophical Association, 46,* 5-20.

Marcus, Ruth B. (1990). Some revisionary proposals about belief and believing. *Philosophy and Phenomenological Research, 50,* 133-153.

Markman, Ellen (1989). *Categorization and naming in children: Problems of induction.* Cambridge, Mass.: MIT.

Martin, C. B. and J. Heil (1996). Rules and powers. Unpublished MS.

McClelland, James L., D. E. Rumelhart, and the PDP Research Group (1986). *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 2.* Cambridge, MA: MIT Press.

McGinn, Colin (1982). *The character of mind.* Oxford: Oxford.

McLaughlin, Brian P. (1988). Exploring the possibility of self-deception in belief. In Brian P. McLaughlin and A.O. Rorty (eds.), *Perspectives on self-deception.* Berkeley, Calif.: University of California.

McMullin, Ernan (1989). The explanation of distant action: Historical notes. In J. T. Cushing and E. McMullin, eds., *Philosophical consequences of quantum theory: Reflections on Bell's theorem.* Notre Dame, Ind.: University of Notre Dame.

Mele, Alfred R. (1987a). *Irrationality: An essay on akrasia, self-deception, and self-control.* New York: Oxford.

Mele, Alfred R. (1987b). Recent work on self-deception. *American Philosophical Quarterly, 24,* 1-17.

Millikan, Ruth (1984). *Language, thought, and other biological categories: New foundations for realism*. Cambridge, Mass.: MIT.

Millikan, Ruth (1993). *White Queen psychology and other essays for Alice.* Cambridge, Mass.: MIT.

Millikan, Ruth (1997). A common structure for concepts of individuals stuffs, and real kinds: More mama, more milk, and more mouse, *Behavioral and Brain Sciences*, forthcoming.

Moore, Chris, C. Gilbert, and F. Sapp (1995). Children's comprehension of the distinction between *want* and *need*. *Journal of Child Language, 22,* 687-701.

Moore, Chris, C. Jarrold, J. Russell, A. Lumb, F. Sapp, F., and F. MacCallum, (1995). Conflicting desire and the child's theory of mind. *Cognitive Development, 10,* 467-482.

Moore, M. Keith, R. Borton, and B.L. Darby (1978). Visual tracking in young infants: Evidence for object indentity or object permanence? *Journal of Experimental and Child Psychology, 25,* 183-198.

Morton, Adam (1980). *Frames of mind.* Oxford: Clarendon.

Mounce, H. (1971). Self-deception. *Proceedings of the Aristotelean Society, 45,* 61-72.

Nagel, Ernest (1979). *The structure of science: Problems in the logic of scientific explanation, 2d ed.* Indianapolis: Hackett.

Nelson, Katherine (1986). *Event knowledge: Structure and Function in Development.* Hillsdale, N.J.: Erlbaum.

Nelson, John O. (1983). Do animals propositionally know? Do they propositionally believe? *American Philosophical Quarterly, 20,* 149-160.

Nisbett, Richard E. and L. Ross (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, N.J.: Prentice-Hall.

Olson, David R. (1988). On the origins of beliefs and other intentional states in children. In J. Astington, P. Harris, and D. Olson (eds.), *Developing Theories of Mind.* Cambridge: Cambridge.

Olson, David R. and R. Campbell (1994). Representation and misrepresentation: On the beginnings of symbolization in young children. In D. Tirosh (ed.), *Implicit and explicit knowledge: An educational approach. Human Development, 6.*

Parkin, Lindsay J. and J. Perner (1997). Wrong directions in children's theory of mind: What it means to understand belief as representation. Unpublished manuscript.

Pears, David (1984). *Motivated irrationality.* Oxford: Clarendon Press.

Perner, Josef (1991a). On representing that: The asymmetry between belief and desire in children's theory of mind. In D. Frye and C. Moore (eds.), *Children's theories of mind.* Hillsdale, N.J.: Erlbaum.

Perner, Josef (1991b). *Understanding the representational mind.* Cambridge, Mass.: MIT.

Perner, Josef (1995). The many faces of belief: Reflections on Fodor's and the child's theory of mind. *Cognition, 57,* 241-269.

Perner, Josef, S. Baker, and D. Hutton (1994). Prelief: The conceptual origins of belief and pretence. In. C. Lewis and P. Mitchell (eds.), *Children's early understanding of mind: Origins and development.* Hove, U.K.: Lawrence Erlbaum.

Perner, Josef, S. R. Leekam, and H. Wimmer (1987). Three-year olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology, 5,* 125-137.

Piaget, Jean (1951). *Play, dreams, and imitation in childhood.* Trans. C. Gattegno and F. M. Hodgson. New York: Norton.

Piaget, Jean (1952). *The origins of intelligence in children.* Tran. Margaret Cook. New York: Norton.

Piaget, Jean (1954). *The construction of reality in the child.* Trans. M. Cook. New York: Basic Books.

Piaget, Jean and B. Inhelder (1969). *The psychology of the the child.* Trans. H. Weaver. New York: Basic Books.

Plato (1961). *Meno*. Trans. W. K. C. Guthrie. In E. Hamilton and H. Cairns (eds.), *The collected dialogues of Plato, including the letters*. Princeton, N.J.: Princeton.

Polanyi, Michael (1969). *Knowing and being.* Ed. M. Grene. London: Routledge.

Posner, George J., K. A. Strike, P. W. Hewson, and W. A. Gertzog, (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education, 66,* 211-227.

Premack, David (1988). 'Does the chimpanzee have a theory of mind' revisited. In R. W. Byrne and A. Whiten (eds.), *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans.* Oxford: Clarendon.

Premack, David and G. Woodruff (1978). Does the chimpanzee have a theory of mind? *The Behavioral and Brain Sciences, 1,* p. 516-526.

Prior, Elizabeth (1985). *Dispositions.* New Jersey: Aberdeen University.

Putnam, Hilary (1962). What theories are not. In E. Nagel, P. Suppes, and A. Tarski (eds.), *Logic, methodology, and philosophy of science: Proceedings of the 1960 international congress.* Stanford: Stanford, 240-251.

Putnam, Hilary (1963). *Brains and behavior.* In R. Butler (ed.), Analytical Philosophy. Oxford: Basil Blackwell.

Putnam, Hilary (1966). The mental life of some machines. In Hector-Neri Castañeda (ed.), *Intentionality, minds, and perception.* Detroit: Wayne State University.

Putnam, Hilary (1975a). The meaning of 'meaning'. In K. Gunderson (ed.), *Minnesota Studies in the Philosophy of Science, 7: Language, Mind, and Knowledge.*

Putnam, Hilary (1975b). *Mind, language, and reality*. New York: Cambridge.

Quine, Willard V. O. (1960). *Word and object.* Cambridge, Mass.: Harvard.

Quine, Willard V. O. (1966/1976). *The Ways of Paradox and Other Essays, rev. ed.* Cambridge, Mass.: Harvard.

Railton (1978/1988). A deductive-nomological model of probabilistic explanation. In J. C. Pitt (ed.), *Theories of explanation.* Oxford: Oxford.

Ramsey, Frank P. (1990). *Philosophical papers.* D. H. Mellor (ed.). Cambridge: Cambridge.

Reddy, Michael (1979). The conduit metaphor. In A. Ortony (ed.), *Metaphor and thought*. Cambridge: Cambridge.

Repacholi, Betty M. and Gopnik, A. (1996). Early reasoning about desires: Evidence from 14- and 18-month-olds. *Developmental Psychology, 33,* 12-21.

Resnick, Lauren B. (1994). Situated rationalism: Biological and social preparation for learning. In L. A. Hirschfeld and S. A. Gelman (eds.), Mapping the mind: Domain specificity in cognition and culture. Cambridge: Cambridge.

Rorty, Amelie O. (1972). Belief and self-deception. *Inquiry, 15,* 387-410.

Rorty, Amelie O. (1988). The deceptive self: Liars, Layers, and Lairs. In Brian P. McLaughlin and A.O. Rorty (eds.), *Perspectives on self-deception.* Berkeley, Calif.: University of California.

Rosch, Eleanor (1977). Human categorization. In N. Warren (ed.), *Advances in cross-cultural psychology, vol. 1.* New York: Academic Press.

Routley, Richard (1981). Alleged problems in attributing beliefs and intentionality to animals. *Inquiry, 24,* 385-417.

Rumelhart, David E., J. L. McClelland, and the PDP Research Group (1986). *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1.* Cambridge, MA: MIT.

Ryle, Gilbert (1949). *The concept of mind.* New York: Barnes and Noble.

Salmon, Wesley C. (1989). Four decades of scientific explanation. In P. Kitcher and W. C. Salmon (eds.), *Minnesota Studies in Philosophy of Science, 13: Scientific explanation.* Minneapolis: University of Minnesota.

Savage, Leonard J. (1972). *The foundations of statistics.* 2d ed. New York: Dover.

Savage-Rumbaugh, E. Sue (1986). *Ape language: From conditioned response to symbol*. New York: Columbia.

Schwitzgebel, Eric (1996). Theories in children and the rest of us. *Philosophy of Science, 63 (proceedings)*, S202-S210.

Searle, John R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3,* 417-424.

Searle, John (1983). *Intentionality*. Cambridge: Cambridge.

Searle, John (1992). *The rediscovery of the mind.* Cambridge, Mass.: MIT.

Searle, John (1994). Animal minds. Unpublished MS.

Shank, Roger C. and R. Abelson (1977). *Scripts, plans, goals and understanding.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Shatz, Marilyn (1994). *A toddler's life: Becoming a person*. New York: Oxford.

Shoemaker, Sydney (1981). Some varieties of functionalism. *Philosophical Topics, 12,* 93-119.

Shultz, Thomas R., D. Wells, and M. Sarda (1980). The development of the ability to distinguish intended actions from mistakes, reflexes, and passive movements. *British Journal of Social and Clinical Psychology, 19,* 301-310.

Sloep, Peter B. and W. van der Steen (1987a). The nature of evolutionary theory: The semantic challenge. *Biology and Philosophy, 2,* 1-15.

Sloep, Peter B. and W. van der Steen (1987b). Syntacticism versus semanticism: Another attempt at dissolution. *Biology and Philosophy, 2,* 33-41.

Smith, Peter (1982). On animal beliefs. *Southern Journal of Philosophy, 20,* 503-512.

Sodian, Beate (1991). The development of deception in young children. *British Journal of Developmental Psychology, 9,* 173-188.

Sodian, Beate, C. Taylor, P. L. Harris, and J. Perner (1991). Early deception and the child's theory of mind: False trails and genuine markers. *Child Development, 62,* 468-483.

Spelke, Elizabeth S., K. Breinlinger, J. Macomber, and K. Jacobsen (1992). Origins of knowledge. *Psychological Review, 99,* 605-632.

Spelke, Elizabeth S., G. Katz, S.E. Purcell, S.M. Ehrlich, and K. Breinlinger (1994). Early knowledge of object motion: Continuity and inertia. *Cognition, 51,* 131-176.

Stalnaker, Robert C. (1984). *Inquiry*. Cambridge, Mass.: MIT Press.

Stampe, Dennis W. (1977). Toward a causal theory of linguistic representation. In P. French et al. (eds.), *Midwest Studies in Philosophy, 2,* 81-102. Minneapolis: University of Minnesota.

Stampe, Dennis W. (1986). Defining desire. In J. Marks (ed.), The ways of desire. Chicago: Precedent.

Stampe, Dennis W. (1987). The authority of desire. *Philosophical Review, 96,* 335-381.

Stampe, Dennis W. (1994). Desire. In S. Guttenplan (ed.), *A companion to the philosophy of mind.* Oxford: Blackwell.

Stich, Stephen (1979). Do animals have beliefs? *Australasian Journal of Philosophy, 57,* 15-28.

Stich, Stephen (1983). *From folk psychology to cognitive science: The case against belief.* Cambridge, Mass.: MIT.

Strawson, Galen (1994). *Mental reality.* Cambridge, Mass.: MIT.

Sullivan, Kate and E. Winner (1991). When 3-year-olds understand ignorance, false belief and representational change. *British Journal of Developmental Psychology, 9,* 159-171.

Sullivan, Kate and E. Winner (1993). Three-year-olds' understanding of mental states: The influence of trickery. *Journal of Experimental Child Psychology, 56,* 135-148.

Suppe, Frederick (1977). *The structure of scientific theories,* 2d ed., Urbana, Ill.: University of Illinois.

Suppe, Frederick (1989). *The semantic conception of theories and scientific realism.* Urbana, Ill.: University of Illinois.

Suppes, Patrick (1962). Models of data. In E. Nagel, P. Suppes, and A. Tarski (eds.), *Logic, methodology, and philosophy of science: Proceedings of the 1960 international congress.* Stanford: Stanford University Press.

Suppes, Patrick (1967). What is a scientific theory? In S. Morgenbesser, ed., *Philosophy of science today.* New York: Meridian.

Thelen, Esther and L. B. Smith (1994). *A dynamic systems approach to the development of cognition and action.* Cambridge, Mass.: MIT.

Thompson, Paul (1983). The structure of evolutionary theory: A semantic approach. *Studies in the History and Philosophy of Science, 14,* 215-229.

Thompson, Paul (1987). A defence of the semantic conception of evolutionary theory. *Biology and Philosophy, 2,* 26-32.

Urmson, J. O. (1956). *Philosophical analysis: Its development between the two world wars.* London: Oxford.

van Fraassen, Bas C. (1970). On the extension of Beth's semantics of physical theories. *Philosophy of Science, 37,* 325-339.

van Fraassen, Bas C. (1972). A formal approach to the philosophy of science. In R. Colodny, ed., *Paradigms and paradoxes.* Pittsburgh: University of Pittsburgh.

van Fraassen, Bas C. (1980). *The scientific image.* Oxford: Clarendon.

van Fraassen, Bas C. (1989a). The charybdis of realism: Epistemological implications of Bell's inequality. In J. T. Cushing and E. McMullin, eds., *Philosophical consequences of quantum theory: Reflections on Bell's theorem.* Notre Dame, Ind.: University of Notre Dame.

van Fraassen, Bas C. (1989b). *Laws and symmetry.* Oxford: Clarendon.

van Fraassen, Bas C. (1991). *Quantum mechanics: An empiricist view.* Oxford: Clarendon.

Vinden, P. 1996: Junín Quechua Children's Understanding of Mind. *Child Development,* 67, 1707-1716.

Vygotsky, Lev S. (1962). *Thought and language.* Ed. and trans. E. Hanfmann and G. Vakar. Cambridge, Mass.: MIT.

Vygotsky, Lev S. (1978). *Mind in society.* Cambridge, Mass.: Harvard.

Weiskrantz, Lawrence (1986). *Blindsight: A case study and implications.* Oxford: Oxford.

Wellman, Henry M. (1990). *The child's theory of mind*. Cambridge, Mass.: MIT Press.

Wellman, Henry M. and M. Banerjee (1991). Mind and emotion: Children's understanding of the emotional consequences of beliefs and desires. *British Journal of Developmental Psychology, 9,* 191-214.

Wellman, Henry M. and K. Bartsch (1988). Young children's reasoning about beliefs. *Cognition, 30,* 239-277.

Willats, John (1984). Getting the drawing to look right as well as to be right: The interaction between production and perception as a mechanism of development. In W. R. Crozier and A. J. Chapman (eds.), *Cognitive processes in the perception of art.* North Holland: Elsevier Science.

Wimmer, Heinz and M. Hartl (1991). Against the Cartesian view on mind: Young children's difficulty with own false beliefs. *British Journal of Developmental Psychology, 9,* 125-138.

Wimmer, Heinz and J. Perner (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in

young children's understanding of deception. *Cognition, 13,* 103-128.

Wimmer, Heinz, G.-J. Hogrefe, and B. Sodian (1988). A second stage in children's conception of mental life: Understanding sources of information. In J. W. Astington, P. L. Harris, and D. R. Olson (eds.), *Developing theories of mind*. New York: Cambridge.

Winner, E. 1982: *Invented Worlds: The Psychology of the Arts.* Cambridge, Mass.: Harvard.

Wittgenstein, Ludwig (1958). *The blue and brown books.* New York: Harper and Row.

Wollheim, Richard (1993). *The mind and its depths.* Cambridge, Mass.: Harvard.

Woodruff, Guy and D. Premack (1979). Intentional communication in the chimpanzee: The development of deception. *Cognition, 7,* 333-362.

Woolley, Jaqueline D. (1995). Young children's understanding of fictional versus epistemic mental representations: Imagination and belief. *Child Development, 66,* 1011-1021.

Wynn, Karen (1992). Addition and subtraction by human infants. *Nature, 358,* 749-750.

Yuill, Nicola (1984). Young children's coordination of motive and outcome in judgments of satisfaction and morality. *British Journal of Developmental Psychology, 2,* 73-81.

Zaitchik, Deborah (1990). When representations conflict with reality: The preschooler's problem with false beliefs and "false" photographs. *Cognition, 35,* 41-68.

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

Zanna, Mark P. and J. Cooper (1974). Dissonance and the pill: An attribution approach to studying the arousal properties of dissonance. *Journal of Personality and Social Psychology, 29,* 703-709.