Philosophy of Science

December, 1996

THE SCIENTIST AS CHILD*

ALISON GOPNIK†‡

Department of Psychology
University of California, Berkeley

This paper argues that there are powerful similarities between cognitive development in children and scientific theory change. These similarities are best explained by postulating an underlying abstract set of rules and representations that underwrite both types of cognitive abilities. In fact, science may be successful largely because it exploits powerful and flexible cognitive devices that were designed by evolution to facilitate learning in young children. Both science and cognitive development involve abstract, coherent systems of entities and rules, theories. In both cases, theories provide predictions, explanations, and interpretations. In both, theories change in characteristic ways in response to counterevidence. These ideas are illustrated by an account of children's developing understanding of the mind.

1. Introduction. Often, progress in science begins with finding the right analogy. Recently, cognitive and developmental psychologists have invoked the analogy of science itself. They talk about our everyday conceptions of the world as implicit or intuitive theories, and about changes in those conceptions as theory changes (Carey 1985, 1988; Karmiloff-Smith 1974, 1988; Gopnik and Wellman 1992, 1994; Gopnik 1984, 1988; Keil 1989; Perner 1991; Wellman 1990; Wellman and Gelman 1992). But to make further progress we need to go beyond analogies: light waves are not wet, planets cannot be made into pies and apples can. We need to specify more substantively and precisely what the real similarities and dif-

*Received November 1995.

†The research and ideas reported in this paper were supported by NSF grant DBS9213959. A portion of it was presented at the Society for Philosophy and Psychology, June, 1995. I am grateful to Henry Wellman, Andrew Meltzoff, Clark Glymour, John Campbell, Philip Kitcher, Eric Schwitzgebel, and two reviewers for illuminating discussions and comments.

‡Send reprint requests to the author, Department of Psychology, University of California, Berkeley, CA 94720.

Philosophy of Science, 63 (December 1996) pp. 485–514. 0031-8248/96/6304-0001$2.00

ferences between children and scientists are, and to show why the similarities matter and the differences do not. We only really make progress when we can specify the common structure that unites light and water, or apples and planets, or children and scientists.

In this article, I will try to give a more substantive and precise characterization of what it means to say that cognitive development is like scientific change. I will describe and defend what some have called "the theory theory." Moreover, I want to argue that the analogy cuts both ways: specifying the parallels between cognitive development and science not only can help us to understand cognitive development, it also can help us to understand science itself. The moral of my story is not that children are little scientists but that scientists are big children. Scientists and children both employ the same particularly powerful and flexible set of cognitive devices. These devices enable scientists and children to develop genuinely new knowledge of the world around them. I will first defend this idea in general terms, then give a more substantive account of what these cognitive devices might be like, and finally give some concrete examples from research on cognitive development.

Many outside the field of developmental psychology (and some inside it) have objected to the scientific analogy on *prima facie* grounds. In fact, the claim that children construct theories is often greeted by scientists, philosophers and psychologists, particularly those with limited experience of anyone younger than a freshman, with shocked incredulity. Surely, they cry, you cannot really mean that mere children construct theories, not real theories, the kind of theories that we, that is we serious grown- up scientists, philosophers and psychologists construct with so much sweat and tears. Aside from injured *amour propre*, these foes of the theory theory point to a number of differences between children and scientists. Scientists are supposed to be consciously—in fact, self-consciously—reflective about their theory-forming and confirming activities. They talk about them and they are part of the scientific stream of consciousness. Only a few adult humans become scientists, there is a division of labor. They do science in a structured institutional setting, in which there is much formal interaction with other scientists. Scientific theory change takes place within the scientific community and a single change may take many years to be completed.

Obviously, none of these things is true of children. Infants and young children do not talk about the fact that they are formulating or evaluating theories, and they certainly do not publish journal articles, present conference papers, or attempt to torpedo the reputations of those who disagree with them. All children develop theories. Conceptual change in children takes place within a single individual and takes place relatively quickly, children may develop and replace many theories in the space of

a few months or years. Children usually converge on the same theories at about the same age. Insofar as these particular types of phenomenology and sociology are an important part of theory formation and change in science, whatever the children are doing is not science.

In order for the theory theory to be more than just a metaphor there has to be some interesting, substantive, cognitive characterization of science, independent of phenomenology and sociology. Is it plausible that science has this kind of cognitive foundation, and that it is similar to the cognitive processes we see in children?

We might imagine that we could turn to the philosophy of science for a simple answer to this question. But philosophers of science have really only begun to consider the question themselves. Historically, the study of science has been divided between normative and sociological approaches to science. One approach details the structure of an ideal scientific inquiry, often as if it involved an abstract set of logical principles. The other focuses on the historical and sociological details of actual scientific practice. Neither the normative nor the sociological projects have had much to say directly about the cognitive foundations of science. The possibility of a cognitive science of science has only just begun to be considered (see e.g Giere 1991).

2. A Cognitive View of Science. What might a cognitive view of science be like? Science is cognitive almost by definition, insofar as cognition is about how minds arrive at veridical conceptions of the world. In one sense scientists must be using some cognitive abilities to produce new scientific theories, and to recognize their truth when they are produced by others. Scientists have the same brains as other human beings, and they use those brains, however assisted by culture, to develop knowledge about the world. Ultimately, the sociology of science must consist of a set of individual decisions by individual humans to produce or accept theories. Scientists eventually converge on the same set of decisions. The view that is the consequence of these decisions converges on the truth about the world. Scientists must be using human cognitive capacities to do this. What else could they be using?

The assumption of cognitive science is that human beings are endowed by evolution with a wide variety of devices that enable us to arrive at a roughly veridical view of the world. Usually in cognitive science we think of these devices in terms of representations and rules that operate on those representations. At any given time, people have some set of representations and rules that operate on them. Over time, there are other cognitive processes that transform both representations and rules. These representations and rules are often an interesting combination of the logical and the psychological: they are abstract structures, often described in terms of

an implicit computational model, but they are also intended to be psychologically real descriptions of how the mind works. The representations and rules may not have any special phenomenological mark, one way or the other, we may not know that we have them, though sometimes we do. They may be deeply influenced by information that comes from other people, but they are not merely conventional and they could function outside of any social community.

We might think of science in terms of such a system of representations and rules. The question that we would ask, then, was whether there were any generalizations to be made about the kinds of representations and rules that underlie scientific knowledge, in particular, and the kinds of processes that transform those representations and rules over time. Is there anything distinctive or special about scientific representations and rules, something that differentiates them from other kinds of representations and rules? Moreover, does the epistemological potency of science, its ability to get things right, come from the nature of these representations and rules, or from some feature of reflective phenomenology or social institutionalization?

A further question, then, would be whether these representations and rules are similar or indeed identical to those we observe in children, and whether changes in those rules and representations over time are like the changes we see in cognitive development. This might be true even if the phenomenology and social organization of knowledge in children and scientists are quite different. And it might be particularly likely to be true if, in fact, the specific phenomenology and sociology of science are not a necessary condition for its epistemological force.

These seem to be straightforward and important questions. It might, of course, turn out that there is, in fact, no distinctive or interesting characterization of the representations and rules that underlie scientific knowledge. Or it might turn out that there is little relationship between the representations and rules of scientists and those of children. Is it worth trying to find out if there is such a relationship?

The detailed empirical work is what I will ultimately turn to, but the project is more plausible and promising than it might seem on the standard view. It is promising not only for our understanding of cognitive development but for our understanding of science itself. A cognitive view of science, and particularly a view that identifies cognitive change in science and childhood, might provide an explanation of the most important thing about science; namely that it gets things right.

Recent work in the study of science presents a dilemma. Science is an activity that is performed by human beings in a social context, and that proceeds in various and haphazard ways. But it nevertheless manifests a kind of logic, and converges on a truthful account of the world. A cog-

nitive view of science might provide an important bridge between normative or logical and sociological views of science. It might indeed be true that there were particular kinds of abstract, logical structures that characterized the important cognitive achievements of science. And it might also be true that looking at the practice of actual science, we might see rather little of that abstract logical structure. But this is quite similar to other cases of cognitive science, from perception to decision-making to parsing to problem-solving. Human beings quite typically acquire knowledge in a way that leads to the truth on average and in the long run, but can also produce errors and incoherencies, that is grounded in a social life but not socially arbitrary. Quite typically, an abstract structure underlies some human cognitive activity that is not at all apparent in superficial phenomenology or practice. Often, that structure is related in interesting ways to the structures we would invent if we constructed an ideal machine to perform that cognitive activity. (We might think of artificial intelligence as a normative enterprise). But that structure is rarely identical to the ideal machine's structure.

A cognitive scientist would say that evolution constructed truth-finding cognitive processes. Science employs a particularly powerful and flexible set of these cognitive abilities. Science uses a set of representations and rules that are particularly well-suited to uncovering the truth about the world. Science gets it right because it uses psychological devices that were designed by evolution precisely to get things right.

The idea that science is related to our ordinary cognition, and that both science and ordinary cognition work for evolutionary reasons is not, of course, new. It is the basic idea behind the "naturalistic epistemology" of Quine and others (Quine 1970; Goldman 1986; Kornblith 1985). I want to propose, however, a more specific version of the naturalistic epistemology story. This view also might be a reason for supposing that the structures of science are particularly likely to be similar to those involved in cognitive development. On this view there might actually be a closer link between science and childhood cognition than between science and our usual adult cognitive endeavors.

Let us go back for a minute to the basic idea that we are endowed by evolution with devices for constructing and manipulating rules and representations and that these devices give us a veridical view of the world. This raises an interesting evolutionary puzzle. Where did the particularly powerful and flexible cognitive devices of science come from? After all, we have only been doing science in an organized way for the last 500 years or so, presumably they did not evolve so that we could do that. I suggest that many of these cognitive devices are involved in the staggering amount of learning that goes on in infancy and childhood. Indeed, we might tell an evolutionary story that these devices evolved, in particular, to allow human children to learn.

Consider a well-established finding in evolutionary psychology. Several distinctive traits are found to correlate with large cortices (relative to body size) across a wide range of species. Moreover, they are also correlated within variants of closely related species. The traits include variation of diet, polygamy, small clutch size, a wide behavioral repertoire, and, most significantly for our case, the relative lack of precocious specialized cognitive abilities in the young, that is, a period of long immaturity (see e.g. Bennett and Harvey 1985).

Passing quickly over polygamy, human beings are, of course, at the extreme end of the distribution on all these other traits and on relative cortical size. From an evolutionary point of view, three of the most distinctive features of human beings are the plasticity of their behavior, their ability to adapt to an extremely wide variety of environments and their long, protected, immaturity. Equipping human children with particularly powerful and flexible cognitive devices, devices that are good at constructing accurate representations of new and unexpected worlds, might be an important part of this evolutionary strategy. We might indeed think of childhood as a period when many of the requirements for survival are suspended, so that children can concentrate on acquiring a veridical picture of the particular physical and social world in which they find themselves. Once they know where they are, as it were, they can figure out what to do. On this view, we might think of infancy as a sort of extended stay in a Center for Advanced Studies, with even better food delivery systems.

It is an interesting empirical question as to how much of this epistemological activity survives in ordinary adult life. Perhaps not much. Once the child has engaged in the theorizing necessary to specify the features of its world, most of us, most of the time, may simply go on to the central evolutionary business of feeding and reproducing. But these powerful theory formation abilities continue to allow all of us at some times, and some of us, namely professional scientists, much of the time, to continue to discover more and more stuff about the world around us. On this view, science is a kind of spandrel, an epiphenomenon of childhood.

2.1 Objection 1: Phenomenology. So I am proposing that the core similarity that we capture in the scientific analogy is a similarity in the rules and representations that allow scientists and children to make cognitive progress. With this cognitive perspective in mind, we can turn back to the phenomenological and sociological differences between scientists and children. Do they undermine the idea that there are deep cognitive similarities between the two groups?

To take the phenomenological question first, it is difficult to see, on the face of it, why conscious phenomenology of a particular kind would play an essential role in finding things out about the world. A characteristic

lesson of the cognitive revolution is that human beings (or, for that matter, machines) can perform extremely complex feats of representation without any phenomenology at all. It is rather characteristic of human cognition that it is largely inaccessible to conscious reflection. Why should this be different in the case of scientific knowledge?

Moreover, the actual degree of conscious reflection in real science is very unclear. It is true that scientists articulate their beliefs about the world or about their fields of scientific endeavor. So, as we will see, do children. But scientists do not typically articulate the processes that generate those beliefs, or that lead them to accept them, nor are they very reliable when they do. The reflective processes are really the result of after-the-fact reconstructions by philosophers of science.

Of course, scientists may sometimes do philosophy of science. They may, from time to time, be reflective about their own activities and try to work out the structure of the largely unconscious processes that actually lead them to form or accept theories. Moreover, there may be circumstances in which this kind of deliberative self-reflection on their own theory formation practices is a real advantage to particular scientists with particular types of problems. However, it seems, at least, much too strong to say that it is a necessary condition for theory formation and change in science. It seems unlikely that it is the reflective phenomenology itself that is what gives scientists their theory formation capacities or that gives those theories their epistemological force.

2.2 Objection 2: Sociology. The sociological objections prompt similar replies. The socially oriented view of philosophy of science has always had a difficult time explaining how science gets it right at all. It has been difficult to reconcile with scientific realism: just as it is hard to see how phenomenology, by itself, could lead to veridicality it is hard to see how a particular social structure, by itself, could do so. Moreover, children are less isolated than the term "little scientist" is likely to imply. They live in a rich social structure with much opportunity for contradiction, instruction, and the linguistic transmission of information. We are not dealing with a contrast between a non-social process and a social one, but between two different types of social organization.

Of course, just as reflective phenomenology might be helpful in solving certain types of problems, so might the characteristic sociological institutions of science. The most striking sociological difference between children and scientists is the division of labor in science, and the resulting complex system of hierarchical social structure. But these features of science seem to have more to do with the kinds of problems children and scientists approach than with the processes they use to solve them. It is characteristic of the child's problems that the evidence necessary to solve

them is very easily and widely available, within crawling distance anyway. It is characteristic of scientific problems that the evidence necessary to solve them is rather difficult to obtain. Formal science quite characteristically applies cognitive processes to things that are too big or too small, too rare or too distant, for normal perception to provide rich evidence. Children, in contrast, typically make up theories about middle-sized, close, perceptible and familiar objects.

It is this paucity of evidence that leads to the division of labor, and to many of the sociological institutions characteristic of science. All you need is a mother and some mixing bowls to find evidence of the spatial properties of objects. To find evidence of Higg's boson you need, quite literally, an Act of Congress. When evidence-gathering becomes this fragmented, complex social relations become more important, and the whole process of finding the truth becomes drastically slower (see Kitcher 1993, for an account of how different social institutions could help scientists to converge on the truth in these circumstances).

It is worth noting that science has become more specialized and institutionalized as the problems of science have become more evidentially intractable. The institutional arrangements of Kepler or Newton or even Darwin were very different from those of contemporary scientists. However, it seems difficult to argue that the basic theory formation capacities of current scientists are strikingly superior to those of Kepler or Newton, in spite of the large differences in social organization.

It is easy to see how the division of labor could result from the need for various kinds of evidence, and how that structure could lead to particular distinctive problems and patterns of timing in scientific change. The social hierarchy and the division of labor, like self-reflection, may be genuinely helpful in solving certain problems. What is extremely hard to see, however, is how the hierarchy could lead to the truth in general, or how the division of labor could itself lead to theory formation or confirmation. The division of labor is one consequence of the different problems children and scientists tackle, and it may be that it gives scientists an advantage in solving those particular problems, just as self-reflection may give scientists an advantage in solving particular problems. However, neither of these facts implies that the basic cognitive resources children and scientists use to tackle those problems are different.

Moreover, in other respects the child's sociological organization may actually be superior to the scientist's for cognitive purposes. Infants and children have infinite leisure, there are no other demands on their time and energy, they are free to explore the cognitive problems that are relevant to them almost all the time. They also have a community of adults who, one way or another, are designed to further the children's cognitive progress (if only to keep them quiet and occupied). Finally, this commu-

nity already holds many of the tenets of the theory the child will converge upon and has an interest in passing on information relevant to the theory to the child.

In fact, we might argue that much of the social structure of science is an attempt to replicate the privileged sociological conditions of infancy. Aside from the division of labor, the social hierarchy largely determines who will get the leisure and equipment to do cognitive work, and to whom other scientists should listen. The infant solves these problems without needing elaborate social arrangements. These are all differences between children and scientists, but again they do not imply differences in the fundamental cognitive processes that the two groups employ.

We might, on my view, think of formal science as a sort of cognitive horticulture. Horticulturalists take basic natural processes of species change, mutation, inheritance and so on, and put them to work to serve very particular cultural and social ends in a very particular cultural and social setting. In the 16th century they bred roses to look like 16th century women, and in the mid-twentieth century they bred roses to look like mid-twentieth century cars. In one sense, an explanation of the genesis of these flowers will involve extraordinarily complex and contingent cultural facts. But in another sense, the basic facts of mutation, inheritance and selection are the same in all these cases, and at a deeper level it is these facts that explain why the flowers have the traits that they do.

In the same way, we can think of organized science as taking natural mechanisms of conceptual change, designed by evolution to facilitate learning in childhood, and putting them to use in a culturally organized way. To explain scientific theory change we may need to talk about culture and society, but we will miss something important if we fail to see the link to natural learning mechanisms.

There is an additional point to this metaphor. Clearly horticulture was, for a long time, the most vivid and immediate example of species change around. And yet, precisely because it was so deeply embedded in cultural and social practices it seemed irrelevant to the scientific project of explaining the origin of species naturalistically. It was only when Darwin, and then Mendel, pointed out the underlying similarities between "artificial" and natural species change that these common natural mechanisms became apparent. Similarly, science has been the most vivid and immediate example of conceptual change around (particularly since most philosophers hang out with scientists more than with children). Its cultural and social features have distracted us from looking at in naturalistic terms. Looking at the similarities between conceptual change in children and in science may yield evidence of a common natural mechanism.

2.3 Objection 3: Timing and Convergence. Another difference we might

point to between children and scientists is that children converge on roughly similar theories at roughly similar times. It might be objected that scientists do not always show this sort of uniform development, and that this weighs against the theory formation view. The theory theory proposes that there are powerful cognitive processes that revise existing theories in response to evidence. If cognitive agents began with the same initial theory, tried to solve the same problems, and were presented with similar patterns of evidence over the same period of time they should, precisely, converge on the same theories at about the same time. These assumptions are very likely to be true for children developing ordinary everyday knowledge. Children will certainly start with the same initial theory and the same theory formation capacities. Moreover, the evidence is ubiquitous and is likely to be very similar for all children.

Notice, however, that, for scientists, these basic assumptions are not usually going to be true. In science, the relevant evidence, far from being ubiquitous, is rare and difficult to come by, and often must be taken on trust from others. The social mechanisms of deference, authority and trust, are, like all social mechanisms, highly variable. Moreover, different scientists also often begin with different theories, and quite typically approach different problems.

In fact, when the assumption of common initial theories and common patterns of evidence, presented in the same sequence, does hold, scientists, like children, do converge on a common account of the world. Indeed, even the timing of scientific discoveries is often strikingly similar, given independent labs working on the same problem with a similar initial theory and similar access to evidence (hence all those shared Nobel Prizes). This convergence to the truth itself is the best reason for thinking that some general cognitive structures are at work in scientific theory change. Scientists working independently converge on similar accounts at similar times, not because evolutionary theory or the calculus or the structure of DNA (to take some famous examples) are innate, but because similar minds approaching similar problems are presented with similar patterns of evidence. The theory theory proposes that the cognitive processes that lead to this convergence in science are also operating in children.

3. What is a Theory? So far I have been trying to argue, in a general way, that scientists and children might employ similar cognitive structures, similar types of rules and representations. Can we characterize these cognitive structures in more detail? What type of rules and representations might be involved in theories and theory change, in both scientists and children? Here is where the metaphor gets to be most useful and interesting for developmental psychologists. Our strategy, at least initially, has largely been simply to adopt the detailed descriptions of theories and theory

change in the philosophy of science, translate them into cognitive psychological terms, and see how well they fit the psychological data.

This approach is quite congruent with some work in the psychology and cognitive science of science. However, it also is differs from the strategies often employed by psychologists of science. Traditionally, psychological studies of science focused on the aspects of science that made it different from ordinary cognition, so that issues of "creativity" or "insight" were at the fore. Conversely, some more recent naturalistic approaches to science have tried to show that scientific knowledge can be reduced to the much broader kinds of knowledge typically discussed in psychology. Theories just are analogies, metaphors, cognitive models, production systems, connectionist nets, scripts, etc. (see, for example, some essays in Giere 1991). The developmental strategy is unlike either of these. Developmental psychologists have found that the typical tools of cognitive science are inadequate for explaining the psychological phenomena we are concerned with. In particular, cognitive science has been notoriously bad at explaining qualitative conceptual change. We want to turn to the example of science for clues to more specific and powerful types of cognitive structures. Theories seem to fit the bill.

One difficulty with this strategy, of course, is that there is much controversy within the philosophy of science about what theories are and how to characterize them. There are the general debates between normative and sociological views. Moreover, within the normative tradition there are extensive debates about the appropriate characterization of the logic of science, most notably a debate between sentential or syntactic and semantic or model theories (van Fraassen, 1980). And there are, of course, debates about the metaphysical status of scientific theories, particularly in domains like quantum physics.

There are interesting possible interactions between these debates and the project of finding a cognitive science of science. Some authors have argued that a model-theoretic view of theories is more compatible with cognitive science than a sentential view. On the other hand a prominent foundational view of cognitive science suggests that all cognitive representations are best understood syntactically (Fodor 1975). Similarly, the naturalistic view seems to presuppose some form of scientific realism, but precisely what form is unclear.

It seems possible, however, to develop a cognitive account of theories that does not necessarily take sides in these debates, and that may be compatible with many different philosophical accounts. We have taken the modest route of focusing on those features of theories that are most generally accepted across many different conceptions of science. Whatever the broader theoretical arguments may be, both normative and sociological investigations of science provide us with rich and suggestive data about

the nature of characteristic scientific representations and rules. So what follows will be brief and, I hope, largely uncontroversial, not to say bland. (For a more extensive version see Gopnik and Wellman 1994; Gopnik and Meltzoff 1997). Even philosophers of science, however, may find that some interesting consequences follow when we consider the truisms of philosophy of science in psychological terms.

Note also that the developmental project is not the typical philosophical project of finding necessary and sufficient conditions by generating counterexamples. Some of what I will say about theories, in general, undoubtedly does not apply to particular instances of scientific theories. As a psychologist I am looking for a natural kind, not a logical one, and I want to use the descriptions of historians and philosophers of science as a starting place for discovering that natural kind.

3.1 Structural Features of Theories. Theories are systems of abstract entities and laws that are related to one another in coherent ways. When we say that theories are abstract, we mean that theoretical constructs are typically phrased in a vocabulary that is different from the vocabulary of the evidence that supports the theory. Theoretical constructs appeal to a set of entities removed from, and underlying, the evidential phenomena they explain.

Theoretical constructs work together in systems with a particular structure. The entities postulated by a theory are closely, "lawfully," interrelated with one another and with the evidence.

Theories also typically appeal to some underlying causal structure that we think is responsible for the superficial regularities in the data. Causal relationships are central to theories in two ways. The intratheoretic relations, the laws, are typically interpreted in causal ways. The mass of an object causes other objects to move towards it, adaptation causes certain mutations to be preserved. But an equally important aspect of theories is that the theoretical entities are seen to be causally responsible for the evidence. The elliptical movements of the planets cause the planets to appear to march across the sky in distinctive ways.

Finally, theories make ontological commitments and support counterfactuals. An accepted theory is supposed to cut nature at its joints, the theoretical entities and laws are supposed to tell you what there is and what it must do. One psychological test of theoreticity is the nature of our surprise at violations of the theory. If we are committed to the theory such violations should strike us, not only as surprising, but as being impossible and unbelievable in an important and strong way. This differentiates theories from other types of knowledge.

3.2 Functional Features of Theories. These structural features of theories

have a number of functional advantages. They allow us to make predictions about new evidence, they help us to interpret evidence and they enable us to explain evidence. A theory, in contrast to a mere empirical generalization, makes predictions about a wide variety of evidence, including evidence that played no role in the theory's initial construction. Some of these predictions will be correct, theories will accurately predict future events described at the evidential level. Others will be incorrect. The ability to produce wide-ranging predictions is perhaps the most obvious pragmatic benefit of science, and it may also be the most important evolutionary benefit of developing theory formation abilities. In fact, making accurate predictions about the behavior of the world and your fellow organisms is the *sine qua non* of cognition.

An additional characteristic of theories is that they produce interpretations of evidence, not simply descriptions and typologies of evidence, and generalizations about it. Indeed, theories strongly influence which pieces of evidence we consider salient or important. From a psychological point of view, theories provide a way of deciding which evidence is relevant to a particular problem, they are a way of solving what computer scientists call "the frame problem". This might also be an evolutionary benefit of theory formation.

A third function of theories that is often mentioned is that they provide explanations. The coherence and abstractness of theories, and their causal attributions and ontological commitments, together give them an explanatory force that mere typologies of the data, or generalizations about it, lack.

In fact, it may be that what we mean by saying that we have explained something is simply that we can give an abstract, coherent, causal account of it. Indeed, it is difficult to find a characterization of "explanation" other than this. From a philosophical point of view, this may be fine. Indeed, often philosophical enterprises consist of giving necessary and sufficient conditions for the application of some term. But from a cognitive point of view, there is a puzzling circularity about most philosophical attempts to explain explanation. The cognitive functions of prediction and interpretation seem obvious enough, representations and rules that allowed for predictions and interpretations would plainly enable an organism to function better in the world. The functional significance of explanation is less clear.

A psychological perspective may help to supply a resolution to this puzzle. Why does it seem to us that explaining is a function of theorizing, but not that theorizing is a function of explaining? The commonsense notion of explanation, at least, seems to involve a kind of affect, a sense of how satisfying a good theory can be. From an evolutionary point, of view we might suggest that explanation is to cognition as orgasm (or at

least male orgasm) is to reproduction. It identifies the motivational and affective state that serves as a marker for the fulfillment of the underlying evolutionary function. An important point of the empirical developmental work, and a common observation about science, is that the search for better theories has a kind of internally-driven motivation, quite separate from the more superficial motivations provided by the sociology. From our point of view, we make theories in search of explanation or make love in search of orgasm. From an evolutionary point of view, however, the relation may be quite the reverse, we search for explanations and orgasms because such a search leads us to make theories and love. We might speculate that we were designed with a theory formation drive and that explanation is a symptom of that drive in action.

3.3 Dynamic Features of Theories. So far we have been talking about the static features of theories, the features that might distinguish them from other cognitive structures such as typologies, schemes, scripts, metaphors, etc. But the dynamic features of theories, the processes involved in theory formation and change, are equally characteristic and perhaps even more important from a developmental point of view. There are characteristic intermediate processes involved in the transition from one theory to another.

The most significant factor in theory change is the accumulation of counterevidence to the theory. However, the initial reaction, as it were, of a theory to counterevidence may be a kind of denial. The interpretive mechanisms of the theory may treat the counterevidence as noise, mess, not worth attending to. At a slightly later stage, the theory may develop ad hoc auxiliary hypotheses designed to account specifically for such counterevidence. Auxiliary hypotheses may also be helpful because they phrase the counterevidence in the accepted vocabulary of the earlier theory. Such auxiliary hypotheses, however, often appear, over time, to undermine the coherence that is one of a theory's strengths.

A next step requires the availability or formulation of some alternative model to the original theory. A theory may limp along for some time under the weight of its auxiliary hypotheses if no alternative way of making progress is available. Often the original idea for the new theory is an extension or application of an idea that is already implicit in some peripheral part of the earlier theory. For example, Darwin takes the idea of selection, which was already widely understood and used in the context of animal breeding and horticulture, and applies it to a new problem. This process may seem like analogy or metaphor, but it involves more serious conceptual changes. It is not simply that the new idea is the old idea applied to a new domain, but that the earlier idea is itself modified to fit its role in the new theory. Moreover, the fertility of the alternative idea

itself may not be recognized immediately. Initially it may only be applied to the problematic cases. We may see only later on that the new idea also provides an explanation for the evidence that was explained by the earlier theory.

A final important dynamic feature of theory formation is the existence of a period of intense experimentation and/or observation. This period might span both the crisis of the earlier theory, the period when anomalous results are being produced right and left, and the first stages of the new theory when a whole range of new predictions become available. The role that experimentation and observation play in theory change is still mysterious, but that it plays some role seems plain. Obviously, experimentation allows the scientist to test the predictions of the theory. But it is worth pointing out that, even in science, much experimentation is much less focused than this, it is more like what we disparagingly call a fishing expedition. In addition to testing predictions, some experimentation enables the scientists to develop a stock of atheoretical empirical generalizations, which in turn will be subsumed and explained by the later theory.

Perhaps the most important dynamic feature of theory change, however, underlying these particulars, is that theory change is, ultimately, caused by evidence. The evidence may take the form of explicit falsifications of the predictions of the theory or it may be the more vague and general result of fishing expeditions. Moreover, the causal sequence by which evidence leads to theory change may be very complex and indirect. From a developmental point of view, this is one of the most important features of theory formation, and the central respect in which it differs from such alternative developmental explanations as maturation and socialization.

4. Theories as Representations. So far I have been borrowing the language of philosophers of science to describe theories and theory change, and I will continue to use that language in describing the children. However, ultimately psychologists want to translate that language into the language of cognitive science, the theoretical parlance of representations and rules. We can think of a theory as a particular kind of system that assigns representations to inputs, in the way that the perceptual system assigns representations to visual input or the syntactic system assigns representations to phonological input. The representations that it assigns are, however, distinctive in many ways, just as perceptual or syntactic representations are distinctive. We can capture these distinctive structural features by talking about the specific abstract, coherent, causal, ontologically-committed, counterfactual supporting entities and laws of the theory, just as we talk about phrase-structures when we describe syntactic representations or $2\frac{1}{2}$ d sketches when we describe perceptual representations. The representations are operated on by rules that lead to new representations, the

theory generates predictions. There are also particular distinctive functional relations between the theoretical representations and the input to them; theories not only predict but also interpret and explain data.

We know that the input to the perceptual and syntactic systems is provided by our sensory systems. Exactly what is the input to the theory system? On one view, we might want to propose that there are other representational systems that translate sensory information into some higher level of ordinary, primary, atheoretical knowledge. On this view, not all our representations would be assigned by theories. Rather an earlier level of processing would provide the evidential input to theory formation processes. Some of these systems might correspond to Fodorian modules. Alternatively, and more in keeping with the philosophical positions that emphasize the "theory-ladeness" of evidence, the system might simply assign theoretical representations to sensory input, without a separate level of evidential representation.

For example, one might make a particular observation, say I see a particular pattern of tracks in a cloud chamber. On any view there will be some very low-level atheoretical perceptual processes that will transform the raw sensory input into some more abstract representational form, say a $2\frac{1}{2}$ d sketch. On the first view we might want to say, however, that there is also a representational system distinct from the theory system which further assigns these inputs a particular "ordinary knowledge" representation. For example, it might represent them as "blue tracks in a white jar on a table." This might then be input to the theory formation system. Alternatively, the theory system might itself simply assign the input a particular theoretical representation, it might just represent the input as electrons decaying in a particular way. For various reasons, this strikes me as a more attractive option than the first one. However, it is an empirical question, and it might differ in different cases.

On both views the theoretical representation assigned to a particular input would then interact in particular rule-governed ways with the other representations of the theory. Does it match the predictions of the theory, for example? Do some particular pair of theoretical representations co-occur in a way that suggests some causal link between them, a link that is not specified in the current theory? In this way, the fact that certain representations occurred and not others might lead to changes within the theory itself. This could happen even if there were no separate evidential level of representation outside the theory itself.

Most significantly and distinctively, however, on any version of the theory theory, the very patterns of representation that occur can alter the nature of the representational system itself. They can alter the nature of the relations between inputs and representations. As we get new inputs, and so new representations, the very rules that connect inputs and repre-

sentations change. Eventually, we may end up with a system with a completely new set of representations and a completely different set of relations between inputs and representations than the system we started out with. Evidence is causally responsible for theory change, and evidence may drive us to construct theories that are radically different from those we initially begin with. This differentiates theories from other kinds of representational systems such as modules, and probably differentiates theoretical representations from perceptual and syntactic ones.

This kind of system may sound so open-ended as to be uninteresting. But in fact the theory formation view proposes that the representational system will change in relatively orderly, predictable, and constrained ways. We can try to capture these features by talking about the dynamic properties of theory change. Factors like counterevidence or the explanatory drive cause the representational system to change, and do so in particular and predictable ways.

Should we think of this relation betwen inputs and outputs as a kind of effective computational procedure? If we know that that the system is in state A and receives input B will we be able to tell what the next set of representations will be like? The answer at this point is simply that we do not know whether there is a procedure like this or not, nor much about how it would operate if there were. What we do know, however, is that there are consistent causal relations between input and representations, both in science and childhood, and my empirical claim is that those relations are similar. The best current explanation of how a brain could instantiate this kind of system of rules and representations is that it is a kind of computer.

There is another, even more profound, question to ask. How does such a system get at the truth about the world? I said before, glibly, that it gets at the truth about the world because it is designed by evolution to get at the truth about the world. But the kind of system I am talking about will certainly suffer from the the same problems of underdetermination that plague the various proposals that were put forward by philosophers of science in the normative tradition. The system will arrive at an answer, given data, and different instantiations of the system will (eventually) arrive at the same answer given the same data. But other answers will still be logically possible given the data.

We might say that the space of relations between the input and output will be very much larger than it will be in a modular system, like the visual perception system, but still smaller than the space of logical possibilities. There will be constraints, though very general constraints, on the kinds of relations between inputs and representations that the system will generate. The constraints correspond to the general assumptions that underly theory formation, that the world has an underlying causal structure, that the

structure is most likely to be the simplest one that corresponds to the data, and so on.

The constraints, on my view, largely come from nature via evolution, and at some level that fact is responsible for their veridicality. Presumably, creatures who constructed representations in different ways in childhood, who did not assume underlying causal structure, did not search for the simplest explanation, did not falsify hypotheses when there was counter-evidence, and so on, were at an evolutionary disadvantage. In that sense nature itself guarantees that the system gets to an understanding of nature.

But, of course, evolution is highly contingent and, for all we know, other systems with different sets of constraints might hit on quite different ways of constructing veridical representations. Perhaps if quantum mechanical effects translated into selection pressures we would have a cognitive system that derived representations from inputs in quite different ways, and we would be less frustrated in our attempts to understand the quantum universe.

So I would say that the system is veridical because of its evolutionary history. However, *how* the process of evolution, or evolution and culture together, actually manages to hit on a system that generates representations that match up to the outside world is still profoundly mysterious.

In any case, however, if are ever to answer these questions, we need first to simply do the descriptive work of specifying what the relations between inputs and representations are like. There is, at least, nothing mystical or incoherent about the idea of a representational system that revises itself in this way. Indeed, we have excellent evidence that just such a system exists in human minds already. Precisely this sort of system generates the representations of science. And we know at least something about how that system characteristically proceeds.

There is an interesting history to this picture of representational change. The picture we are presenting here actually has many features in common with the picture of science (and of knowledge more generally) that emerged in the logical empiricist tradition in the late fifties. The early empiricist tradition, the tradition we might call positivism, had attempted to translate or define the terms of scientific theories (and other kinds of knowledge) into some vocabulary of primitive perceptual terms, "sense-data." The project was actually similar in many respects to the projects of much contemporary empiricist theorizing in cognitive science and psychology. By the fifties the difficulties of that project had become quite clear. One reaction—the reaction that prevailed in "the cognitive revolution," particularly in the work of Chomsky—was, quite explicitly, to return to the apriorist rationalist tradition of Kant and Descartes.

But within the tradition itself the reaction took a rather different form. Carnap's late work and Quine's early work both articulated a view in

which the project was not any longer to redefine the theories of science in a primitive perceptual vocabulary. Rather one could see the progress of science (and knowledge more generally) as the successive articulation and replacement of a series of theories (Carnap calls them languages). The first in the series of theories is what Carnap calls the thing-language, and what we might think of as the language of everyday "folk physics." For Carnap, there are systematic relations between the statements of the earlier language and those of the later language, but these relations are much weaker than definition. There are many logically possibly alternative languages that could be constructed that could still have the appropriate relation to the earlier language; one's choice among them will be guided by a number of practical scientific considerations (never made entirely clear). One might say that Quine translated Carnap's formal account into a memorable picture in "Two Dogmas of Empiricism", the picture of the web of belief. (Pictures speaking louder than logical symbols for most of us, Quine got the credit).

The worm in the apple, the serpent in the garden, of these accounts was the idea, which is in both Quine and Carnap, that the choice of the new language was "conventional" or "pragmatic." The implication, later made quite explicit in Quine, was that the decision was also arbitrary and simply socially determined. If Skinner's "Verbal Behavior" informed your view of how new languages were constructed, then this arbitrary, conventionalist view made sense. This idea set the stage for the later skepticism of Quine and the social constructivists who were to follow him.

The "theory theory" view of the development of knowledge would adopt the picture of a succession of theories but would replace convention with nature. The choice of which theory to move to, which language to make up next, is not simply arbitrary or conventional. Rather it is the result of the operation of psychological devices designed by evolution to lead to veridical outcomes. Moreover, the initial theory, the first language, is not the "thing-language," but can be determined empirically by looking at the initial conceptions of infants.

5. Theories in Childhood. I want to claim that infants and young children have cognitive structures that are like those we have just been describing. Children's early cognitive structures should have these characteristics if they are really theoretical. That is, they should involve appeals to abstract theoretical entities, with coherent causal relations among them. Theories should lead to characteristic patterns of predictions, including extensions to new types of evidence and false predictions, not just to more empirically accurate predictions. Theories should also lead to distinctive interpretations of evidence, a child with one theory should interpret experiences differently than a child with a different theory. Finally, theories should

invoke characteristic explanations phrased in terms of these abstract entities and laws. This distinctive pattern of prediction, interpretation, and explanation is among the best indicators of a theoretical structure and the best ways of distinguishing the theory theory from such developmental competitors as scripts, schemas, and simulations (see Gopnik and Wellman 1992).

The dynamic features I have described should be apparent in children's transitions from one theory to a later one. Children should ignore certain kinds of counterevidence initially, then account for them by auxiliary hypotheses, then use the new theoretical idea in limited contexts, and only finally reorganize their knowledge so that new theoretical entities play a central role. During the period when the new theory is, as it were under construction, they should engage in extensive experiments relevant to the theory, and collect empirical generalizations. Over a given developmental period, we should be able to chart the emergence of the new consistent theory from the earlier one, and we should be able to predict a period of some disorganization in between. Moreover, children should construct different theories if they receive different patterns of evidence. These dynamic features of theories also help to distinguish them from other types of cognitive structures. In particular, they make theories different from modules (see Gopnik and Wellman 1994; Gopnik 1995; Gopnik and Meltzoff 1997).

If, in fact, we discovered that children's representations and rules were like this, we would be licensed in saying that children had theories, and that the process of theory construction was similar in scientists and children.

5.1 An Example—The Child's Changing Theories of the Mind. Well, are children's representations and rules like this? Even if the theory theory is plausible, is it true? The best argument for any empirical claim is, of course, the data. The enthusiasm for the scientific metaphor in psychology has stemmed from its real advantages in explaining psychological phenomena. Two lines of development have been particularly important. First, psychologists have looked at our everyday categorizations of objects. Earlier theories tried to explain these categorizations in terms of the perceptual features of the objects. More recent research suggests that categorization is best understood in terms of our everyday theories about the underlying causal structure of objects (Murphy and Medin 1985). If we look at young children's categorization, both in language and behavior, we see a very similar pattern. Even two- and three-year-old children appear to categorize objects in terms of "natural kinds," underlying essences with causal efficacy. Moreover, their decisions about which objects belong to these natural kinds appear to be rooted in naive theories of physics and

biology. These very young children have coherent abstract accounts of objects and animals and use those accounts to generate predictions and explanations (Carey 1985; Keil 1989; Gelman 1986; Gelman and Wellman 1991). Most significantly, it is possible to chart qualitative conceptual changes in children's categorization as their theories are constructed, modified, and revised. Thus, in Carey's work, for example, the child categorization of an object as an "animal" or as "alive" changes profoundly as the child's "folk biology" changes (Carey 1985).

Second, the renewed interest in "folk psychology" has raised the possibility, first formulated in the philosophical literature, (Churchland 1984, Stich 1983) that our everyday understanding of the mind is analogous to a scientific theory. Again, empirical investigations of the child's developing understanding of the mind have tended to confirm this view. The majority of investigators in the field have argued that the child's early understanding of the mind can be usefully construed as a theory and that changes in that understanding can be thought of as theory changes (Gopnik 1993; Gopnik and Wellman 1994; Wellman 1990; Perner 1991, Flavell et al. 1995, though see Harris 1991 and Leslie 1991 for opposing views). In the course of developing an account of the mind children postulate such mental entities as perceptions, beliefs and desires as a way of explaining ordinary human action. Moreover, there are significant and far-reaching conceptual changes in the child's understanding of the mind, much like theory changes.

In the remainder of this paper, I will briefly outline some examples of the way that we have used the analogy to science to understand the child's developing understanding of the mind. I will not have the space to go into all the empirical details or to consider all the possible alternative explanations, the interested reader can turn to the empirical reports. Instead, I will simply give a brief sketch of part of the developmental story. I will work backwards from 4-year-olds to newborns, suggesting at each point that the child's knowledge is theoretical and differs from earlier knowledge in a way that suggests theory change.

Four-year-old children understand a surprising amount about the nature of the mind. They make consistent and largely, though not invariably, correct predictions about a wide variety of events, including new events quite different from events they have previously experienced. We can present these children with a hypothetical event, for example, a story about another child who is deceived. This hypothetical child is presented with a candy box that is really full of pencils, or a "Hollywood rock" that is really made of sponge, or he sees his mother put chocolate in the green cupboard and the chocolate is surreptitiously switched to the blue cupboard while he is out of the room. Many other stories have also been explored. Then we can ask the children to make a wide variety of predic-

tions about future events. Where will the other child look? What will he think? What will he say? How will he feel? How will the object look to him? We can even present these questions as counterfactuals. How would he feel if it were otherwise? What would you say if you were him? Four- and five-year-olds make all these predictions accurately (Gopnik and Astington 1988, Flavell 1986, Perner 1991, Wellman 1990; Wimmer and Perner 1983).

Moreover, these children justify their predictions by offering causal explanations. The explanations relate the evidence of action to an underlying apparatus of beliefs and desires, mental entities and psychological laws, and they relate different mental entities to one another in a coherent way. Why did he look in the cupboard? He looked in the cupboard because he wanted the chocolate and thought it was in the cupboard. How did he know it was in the cupboard? He knew it was in the cupboard because he saw his mother put it there. The children do this both in prompted experimental situations and in their natural spontaneous language (Bartsch and Wellman 1989, 1995; Wellman 1990).

They also interpret action in terms of these underlying psychological entities. Presented with a neutral description of human behavior they automatically interpret and describe it in terms of beliefs and desires (Lillard and Flavell 1990). And they make quite explicit ontological claims about mental events. In particular, they make an ontological distinction between mental and physical objects (actually they seem to be committed to a kind of substance dualism) (Wellman and Estes 1986). So it appears that the four-year-old's understanding of the mind has some of the structural and functional character of a theory. It postulates abstract entities, beliefs, and desires, which are coherently interrelated, with consistent causal relations to each other and to the evidence of action. The children make ontological commitments and causal attributions. They predict, explain and interpret.

Does the children's knowledge have the dynamic character of a theory, however? In particular, is it constructed from an earlier theory or is it discovered through introspection, or does it mature or become internalized in social interaction? To answer that we must turn to even younger children, 2½- and 3-year-olds. Three-year-olds also generate predictions about the behavior and mental states of others, they make interpretations and provide explanations. But those predictions, interpretations and explanations are quite different from those of the older children.

Three- and even 2½-year-olds make strikingly wide-ranging and accurate predictions about some mental states, particularly perceptions and desires. They predict accurately, for example, that another person on the other side of a screen will not be able to see what they see themselves (Flavell et al. 1981, Masangkay et al. 1974). They predict that those with different desires will perform different actions, and be made happy or sad

by different things (Astington and Gopnik 1991; Wooley and Wellman 1990, Yuill 1984). Even these very young children will also give coherent causal explanations of actions in terms of perception and desire. Again these predictions are manifested both in experiments and in spontaneous speech and behavior (Wooley and Wellman 1990; Bartsch and Wellman 1989, 1995).

But when we turn to the young three-year-old's understanding of belief we see quite a different pattern. Three-year-olds consistently predict that the children in the deceptive cases will act on the reality and not on their false belief. They inaccurately but consistently predict that the child will think there are pencils in the box or that the chocolate is in the green cupboard, or that the rock is a sponge. They make similarly inaccurate predictions about emotions and actions, the child will look for the chocolate in the right cupboard. She will not be disappointed when she sees the pencils (Gopnik and Astington 1988, Perner 1991, Wellman 1990, Wimmer and Perner 1983).

More strikingly, their explanations, both in spontaneous speech and in experimental situations are consistent with these predictions. For example, they give an inadequate causal account of the sources of beliefs, they are unable to discriminate, between what can be learned by seeing something and what can be learned by drawing an inference about it (Gopnik and Graf 1988; O'Neill, Astington, and Flavell 1992). This kind of rich causal account of the source of beliefs is necessary to support correct predictions about false beliefs. Two-year-olds explain actions in terms of desires and, as the third year progresses, in terms of true beliefs, but not in terms of false beliefs (Bartsch and Wellman 1995). Most strikingly of all, they misinterpret and misreport the very data that falsify their predictions, and confirm the four-year-old's theory. Suppose we present them with a child actually opening the deceptive box and manifesting surprise (Moses and Flavell 1990). Or even better suppose we actually let them experience the deception themselves (Gopnik and Astington 1988, Gopnik 1993). Young three-year-olds in these circumstances simply misinterpret the data. They insist, in spite of the data, that both the other and they themselves always knew, and always acted on, the truth (notice that their inaccuracy in their own case rules out the possibility that they learn about the mind through some sort of privileged first-person introspective knowledge; see Gopnik 1993).

These phenomena have suggested to a number of investigators in the field that the difference between the three- and four-year-olds is best characterized as a difference in the implicit theories of the world that the children hold, though there is some debate about exactly how to characterize that theoretical difference. Broadly, 2½- and young three-year-olds seem to understand the mind largely in terms of desires and perceptions, and

they think perception is always veridical. As long as you are, as it were, perceptually pointed at the right part of the world you will know everything there is to know about it. They make incorrect predictions or explanations in cases where the causal relation between the mind and world is more complex, such as cases of false belief.

The differences between three- and four-year-olds are now well-established though there are, of course, debates about the best way to explain them. More recently, there are starting to be some hints about the transition from the earlier view to the later one. Older three-year-olds often show an interesting transitional pattern, in which they will occasionally make the correct predictions about the belief in limited circumstances. One relevant circumstance seem to be when they are forced to confront and explain counterevidence to the theory (Mitchell and LaCohee 1991, Wellman and Bartsch 1988). Another is when the belief problems are placed in the context of the child's well-established understanding of desire or perception (Gopnik and Slaughter 1991, Gopnik et al. 1994, Flavell and Moses, 1994, Moses 1993). These children seem to show the first glimmers of the new theory of belief when it must be recruited in special circumstances to explain counterevidence. It initially functions as a kind of auxiliary hypothesis. And they also seem to show the first glimmers of the new theory when the analogies between problems of belief and problems of desire and perception are made particularly clear and salient. Initially, however, they only use the idea of belief in general, and false belief, in particular, in these limited contexts. They fail to apply them widely in an explanatory and predictive way.

Moreover, three-year-old children show some signs of intense experiment and observation during this transition. They become fascinated by cases of deception or misleading appearance, or of differences in their own mental states and those of others, and they explore the nature of those differences (Bartsch and Wellman 1995).

There are also even more interesting but even more tentative hints emerging in the literature that the transition from one theory to another is actually caused by the accumulation of evidence and counterevidence. The first hint is that children with many siblings consistently make the theory change earlier than children with fewer siblings. (This is in spite of the fact that children with fewer siblings do better on measures of general cognitive achievement like IQ; see Perner and Ruffman 1992.) The most common explanation that has been proposed for this finding is that siblings provide a particularly rich source of evidence about the diversity of beliefs and desires. Similarly, children appear to make the theory change sooner if their families consistently discuss mental states at the dinner table, another important source of evidence (Dunn et al. 1991).

Most concretely and convincingly, we can actually reliably induce the

theory change in 3½-year-olds. Slaughter did this by providing the children with relevant counter evidence over a two-week period in a training study. Interestingly, evidence about perception and desire, as well as belief was effective in improving the child's understanding of belief. Moreover, children who received salient counterevidence also showed improved understanding of other parts of the theory that are coherently linked to an understanding of belief, such as an understanding of the sources of information (Gopnik et al. 1994, Slaughter and Gopnik in press).

These sorts of investigations of the dynamic aspects of theories in developmental psychology may ultimately have the most to contribute to our understanding of science. Theory change proceeds more uniformly and quickly in children than in scientists, and so is considerably easier to observe, and we can even experimentally determine what kinds of evidence lead to change. In children, we may actually be able to see "the logic of discovery" in action.

Can we trace the theory formation process back even further? The lower limit for purely linguistic tasks like the ones we have described so far is about 2½-years-old. With some ingenuity, however, we can use other kinds of paradigms to explore even younger children's predictions about the mind. We mentioned that by 36 months or so children show a secure understanding of the diversity of perceptions and desires. Some recent work in our laboratory suggests that even this understanding is constructed by revising an earlier incorrect theory. We devised a task to explore 24- and 30-month-old's predictions about perception. The youngest children consistently made incorrect predictions, when they were instructed to hide the object from the experimenter they responded by placing the object on the experimenter's side of the screen, so that it was invisible to them and visible to the experimenter. In contrast, most of the 36-month-olds made the correct predictions both behaviorally and verbally. Most interestingly, the 24- and 30-month-olds engaged in extensive experimental behaviors in these settings. They moved objects from one side of the screen to the other and frequently got up themselves and moved to the other side of the table to examine the result (Gopnik et al. 1994; see also Lempers et al. 1977)

Repacoli designed a similar experiment examining the origins of an understanding of the diversity of desire. The experimenter tasted goldfish crackers and made a disgusted face and tasted raw broccoli and made a delighted face. She then held out her hand to the child and said "give me some." 18-month-olds correctly inferred her desire: they gave her broccoli. In contrast, 14-month-olds made a relevantly incorrect prediction: they gave her goldfish crackers. Just as in the case of understanding perception, the incorrect predictions of the earlier theory set the stage for the advances of the later theory (Repacoli and Gopnik 1995).

Also, as in that case, there is some evidence of experimental behavior at this point. One interpretation of the apparently irrational behavior of the "terrible twos" is that it represents a series of experiments exploring the divergences between the child's desires and those of others. (Indeed it is the very cold-bloodedness with which a toddler slowly reaches for the object you have expressly forbidden, carefully examining your face all the time, that makes these behaviors so infuriating. If the child is a psychologist, then we parents are the laboratory rats).

Of course, as we deal with younger and younger children, it becomes more difficult to demonstrate that their understanding is genuinely theoretical. We have evidence of consistent patterns of prediction and interpretation, and of experimentation, but not of explicit explanation and ontological commitment. Ideally we would want to show that these predictions are consistent, wide-ranging and productive, and that these experiments are systematic and are related to the cognitive changes. Moreover, we would want to show that new evidence is causally responsible for the changes in predictions and interpretations. (We have a bit more substantial evidence along these lines for other theoretical changes in infancy, particularly changes in infants' understanding of objects; see Gopnik 1988, Gopnik and Meltzoff 1997.) It is striking, however, that the child's knowledge at each stage seems conceptually related to knowledge at the succeeding stage. An incorrect view of perception leads to incorrect predictions about action; this is replaced by a better view of perception that leads to incorrect predictions about false beliefs, and so on.

At this point some of the psychologists, scientists and philosophers who were crying out at the start of this paper, may well be doing so again; "Surely, you cannot think it is theories all the way down!" Well, yes, actually, I do think it is theories all the way down. Andrew Meltzoff and I have argued that, at birth, infants already draw some inferences about human behavior that go well beyond the direct evidence of their senses (Gopnik and Meltzoff 1994, 1997, Meltzoff and Gopnik 1993). In particular, young infants already seem to make rather abstract mappings between the bodily movements of other people and their own internal states and to draw at least a primitive kind of inference and prediction on this basis. These inferences are apparent in infants' early imitation of facial gestures and in their more complex interactions with other people. Infants seem to have innate knowledge of the mind, and this knowledge is theory-like, at least in the sense that it goes well beyond immediate perceptual experience, that it enables genuine and productive predictions, and that it is revised in the light of further evidence. (When we say that this knowledge is innate we do not mean this in the philosophical sense, which is that neither the philosopher in question nor any of the guys down the hall could think of a way to learn it. We mean that it has been demonstrated in 42-minute-old infants (Meltzoff and Moore 1983).)

Notice that this innate theory gives human beings a tremendous boost in solving the apparently intractable problem of inferring the minds of others. It is possible that the severe difficulties of people with autism, for example, stem from the absence of this initial rich mapping from the self to other people (Meltzoff and Gopnik 1993). At the same time, however, the initial theory sets us up for exactly the problems that later theory formation processes must solve. We innately assume that our mental states are the same as those of others and only an extended process of evidence gathering, experimentation, and theoretical revision allows us to give a proper explanation of the many cases in which they are not.

6. Conclusion. At this point, the analogy between science and development has been of most benefit to developmental psychologists. In the particular case of the child's understanding of the mind, it has also been illuminating to philosophers of mind. Apparently philosophical questions about the nature of our adult knowledge of the mind can be answered by developmental evidence. For example, a theory view and an introspectivist or simulation view of adult knowledge lead to quite different predictions about development, predictions that can be empirically tested (see essays in Davies and Stone 1995).

If the analogy is correct, however, it should be important and interesting to philosophers of science as well. In particular, any account of science will, at the least, need to be able to explain the apparently striking similarities between the two domains. Accounts that stress the sociology of science, for example, will have to explain why children's cognitive processes are apparently so similar to scientists' when their sociology is quite different. More positively, the analogy suggests a naturalistic answer to one of the most puzzling and important questions about science: why it is that we human beings should be able to discover such peculiar things about the universe we live in. So far the details of the developmental answer to that question have largely been borrowed from classical accounts of scientific change. But we might hope that the borrowing will eventually go in both directions, that, for example, studying the transitional processes in the three- to four-year-old's theory of mind might illuminate historical processes of theory change in science. Finally, the analogy also might contribute to normative investigations of science. Some philosophers might say that they do not want to know how mere contingent human brains happen to learn about the world, but rather how it is that any device could possibly learn about the world. But, if we are serious about that project, we should also want to know how evolution actually did construct the device that is the best learner on Earth: the human child.

REFERENCES

Astington, J. 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, pp. 39–50.

Bartsch, K. and H. M.Wellman (1989), "Young Children's Attribution of Action to Beliefs and Desires", *Child Development* 60(4): 946–964.

———. (1995). *Children Talk about the Mind*. New York: Oxford University Press.

Bennett, K. and P. Harvey (1985), "Brain Size, Development and Metabolism in Birds and Mammals", *Journal of Zoology* 207: 491–509.

Carey, S. (1985), *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.

———. (1988), "Conceptual Differences between Children and Adults", *Mind and Language* 3(3): 167–183.

Churchland, P. (1984), *Matter and Consciousness*. Cambridge, MA: Bradford/MIT Press.

Davies, M. and T. Stone (ed.) (1995). *Folk psychology: The Theory of Mind Debate*. Oxford: Basil Blackwell.

Dunn, J., J. Brown, C. Slomkowski, C Tesla, and L. Youngblade (1991), "Young Children's Understanding of Other People's Feelings and Beliefs: Individual Differences and Their Antecedents", *Child Development* 62: 1352–1366.

Flavell, J. H., B. A. Everett, K. Croft, and E. R. Flavell (1981), "Young Children's Knowledge about Visual Perception: Further Evidence for the Level 1–Level 2 Distinction", *Developmental Psychology* 17: 99–103.

Flavell, J. 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(4): 915–928.

Flavell, J. 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, 51, No. 1.

———. (1995), *Young Children's Knowledge about Thinking*. Monographs of the Society for Research in Child Development.

Gelman, S. A. and H. M. Wellman (1991), "Insides and Essence: Early Understandings of the Non-Obvious", *Cognition*: 213–244.

Giere, R. (ed.) (1991), *Cognitive Models of Science*. Minneapolis: University of Minnesota Press.

Goldman, A. (1986), *Epistemology and Cognition*. Cambridge, MA: Harvard University Press.

Gopnik, A. (1984), "Conceptual and Semantic Change in Scientists and Children: Why There Are No Semantic Universals", *Linguistics* 20: 163–179.

———. (1988), "Conceptual and Semantic Development as Theory Change", *Mind and Language* 3:197–217.

———. (1993), "How We Know Our Minds: The Illusion of First-Person Knowledge of Intentionality", *Behavioral and Brain Sciences* 16: 1–14.

———. (1996), "Theories and Modules: Creation Myths, Developmental Realities and Neurath's Boat", in P. Carruthers and P. Smith (ed.), *Theories of Theory of Mind*. Cambridge: Cambridge University Press, pp. 169–183.

Gopnik, A. and J. W. Astington (1988), "Children's Understanding of Representational Change and its Relation to the Understanding of False Belief and the Appearance-Reality Distinction", *Child Development* 59: 26–37.

Gopnik, A. and P. Graf (1988), "Knowing How You Know: Young Children's Ability to Identify and Remember the Sources of Their Beliefs", *Child Development* 59: 1366–1371.

Gopnik, A., and A. N. Meltzoff (1994), "Minds, Bodies and Persons: Young Children's Understanding of the Self and Others as Reflected in Imitation and "Theory of Mind" Research", in S. Parker and R. Mitchell (ed.), *Self-Awareness in Animals and Humans*. New York: Cambridge University Press, pp. 157–181.

———. (1997), *Words, Thoughts, and Theories*. Cambridge, MA: Bradford/ M.I.T. Press.

Gopnik, A., A. N. Meltzoff, and V. Slaughter (1994), "Changing Your Views: How Understanding Visual Perception Can Lead to a New Theory of the Mind", in C. Lewis and P. Mitchell (ed.) *Origins of a Theory of Mind*. New Jersey: Erlbaum.

Gopnik, A. and V. Slaughter (1991), "Children's Understanding of Changes in Their Mental States", *Child Development* 62: 98–110.

Gopnik, A. and H. Wellman (1992), "Why the Child's Theory of Mind Really is a Theory", *Mind and Language* 7(1 and 2): 145–172.

———. (1994), "The 'Theory Theory'", in L. Hirschfield and S. Gelman (ed.), *Mapping the Mind: Domain Specificity in Culture and Cognition*. New York: Cambridge University Press, pp. 257–293

Harris, P. (1991), "The Work of the Imagination", in A. Whiten (ed.), *Natural Theories of Mind: The Evolution, Development, and Simulation of Second-Order Mental Representations*. Oxford: Basil Blackwell, pp. 283–304.

Karmiloff-Smith, A. (1988), "The Child is a Theoretician, Not an Inductivist", *Mind and Language* 3(3): 183–197.

Karmiloff-Smith, A. and B. Inhelder (1974), "If You Want to Get Ahead, Get a Theory", *Cognition* 3(3): 195–212.

Keil, F. (1987), "Conceptual Development and Category Structure", in U. Neisser (ed.), *Concepts and Conceptual Development*. New York: Cambridge University Press, pp. 175–201.

Keil, F. C. (1989), *Concepts, Kinds, and Cognitive Development*. Cambridge, MA: MIT Press.

Kitcher, P. (1993), *The Advancement of Science*. Oxford: Oxford University Press.

Kornblith, H. (ed.) (1985), *Naturalizing Epistemology*. Cambridge, MA: MIT Press.

Lempers, J. D., E. R. Flavell, and J. H. Flavell (1977), "The Development in Very Young Children of Tacit Knowledge Concerning Visual Perception", *Genetic Psychology Monographs* 95: 3–53.

Leslie, A. M. (1991), "Information Processing and Conceptual Knowledge: The Theory of TOMM", paper presented at the meetings of the Society for Research in Child Development, Seattle, WA.

Lillard, A., and J. Flavell (1990), "Young Children's Preference for Mental State Versus Behavioral Descriptions of Human Action", *Child Development* 61(5): 731–741.

Masangkay, Z., K. McCluskey, C. McIntyre, J. Sims-Knight, B. Vaughan, and J. H. Flavell (1974), The Early Development of Inferences About the Visual Percepts of Others", *Child Development* 45: 357–366.

Meltzoff, A. N. and A. Gopnik (1993), "The Role of Imitation in Understanding Persons and Developing a Theory of Mind", in S. Baron-Cohen, H. Tager-Flusberg, and D. Cohen (ed.), *Understanding Other Minds: Perspectives From Autism*. Oxford: Oxford University Press, pp. 335–366.

Meltzoff, A. N. and M. K. Moore (1983), "Newborn Infants Imitate Adult Facial Gestures", *Child Development* 54, 702–709.

Mitchell, P. and H. Lacohee, H. (1991), "Children's Early Understanding Of False Belief", *Cognition* 39(2): 107–109.

Moses, L. J. (1993), "Young Children's Understanding of Belief Constraints on Intention", *Cognitive Development*, pp. 1–27.

Moses, L. J., and J. H. Flavell (1990). "Inferring False Beliefs from Actions and Reactions. *Child Development* 61(4): 929–945.

Murphy, G. and D. Medin (1985), "The Role of Theories in Conceptual Coherence", *Psychological Review* 92: 289–316.

O'Neill, D. K., J. W. Astington, and J. H. Flavell (1992), "Young Children's Understanding of the Role that Sensory Experiences Play in Knowledge Acquisition", *Child Development* 63(2): 474–491.

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

Perner, J., T. Ruffman, and S. R. Leekam (1994), "Theory of Mind is Contagious: You Catch It from Your Sibs", *Child Development* 65, 5: 1228–1238.

Quine, W. V. O., and J. S. Ullian (1970), *The Web of Belief*. New York: Random House.

Repacoli, B. and A. Gopnik (in press), "Early Reasoning about Desires: Evidence from 14- and 18-Month-Olds", *Developmental Psychology*.

Slaughter, V. and A. Gopnik (in press), "Conceptual Coherence in the Child's Theory of Mind", *Child Development*.

Stich, S. (1983), *From Folk Psychology to Cognitive Science: The Case Against Belief*. Cambridge, MA: Bradford Books/MIT Press.
Van Fraassen, B. (1980), *The Scientific Image*. Oxford; Oxford University Press.
Wellman, H. (1990), *The Child's Theory of Mind*. Cambridge, MA: Bradford Books/MIT Press.
Wellman, H. M. and K. Bartsch (1988), "Young Children's Reasoning about Beliefs", *Cognition* 30: 239–277.
Wellman, H. M. and D. Estes (1986), "Early Understanding of Mental Entities: A Reexamination of Childhood Realism", *Child Development* 57: 910–923.
Wellman, H. M. and S. A. Gelman (1992), "Cognitive Development: Foundational Theories of Core Domains", *Annual Review of Psychology* 43: 337–375.
Wellman, H. M. and J. D. Woolley (1990), "From Simple Desires to Ordinary Beliefs: The Early Development of Everyday Psychology", *Cognition* 35(3): 245–275.
Wimmer, H., and J. Perner (1983), "Beliefs About Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception", *Cognition* 13: 103–108.
Yuill, N. (1984), "Young Children's Coordination of Motive and Outcome in Judgements of Satisfaction and Morality", *British Journal of Developmental Psychology* 2: 73–81.