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, ... O_n} be the set of objects in the environment that are possible referents of W. Let {V_1, V_2, ... 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}, ... 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.) <u>Abstractness</u>. 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.