8 The Emergence of Theoretical Beliefs as Constraints on Concepts

Frank C. Keil
Cornell University

Cognitive psychology has recently embraced a view of concepts that has had a long tradition in the philosophy of science, namely that coherent sets of core beliefs, or "theories," are essential to a full specification of concept structure. Concepts cannot be represented merely in terms of probabilistic distributions of features or as passive reflections of feature frequencies and correlations in the world. Some of the most compelling demonstrations involve illusory correlations where prior theories cause people to create or enhance correlations that are central to their theories and ignore or discount equally strong correlations that are more peripheral to that theory. This phenomenon has been known for some time in the social and clinical psychology literature, such as in the illusory correlations in diagnoses made by clinical psychologists (e.g., Chapman & Chapman, 1969); but its greater relevance to most concepts is now being widely recognized (Murphy & Medin, 1985).

There are many other problems with mere probabilistic models, such as demonstrations that equally typical (i.e., equally probabilistically associated) features may be dramatically different in how they affect judgments about the goodness of exemplars. Thus, Medin and Shoben (1988) have shown that, although curvedness is judged to be equally typical of bananas and boomerangs, straight boomerangs are considered to be much more anomalous members of the boomerang family than straight bananas in their family, because curvedness is seen as theoretically more central, that is, causally more critical to the "essence" of boomerangs. This finding is also further evidence against real-world correlations exclusively driving concept structure because, empirically, there are, in fact, some straight boomerangs and no straight bananas.

Still other examples involve older demonstrations by Asch (1952) of the extent to which the features that make up concepts of persons are heavily interactive. If an unknown person is described with a list of six traits, such as: intelligent, skillful, industrious, warm, determined, practical, and cautious, a certain impression will form. Changing the value of one feature, such as warm to cold, creates a different overall impression through interactions with many other features. Thus, one cannot usually change one feature and expect the effect to be limited to that feature. These interactions may be best understood in terms of subjects' possession of an implicit theory of causal factors responsible for the emergence of behavioral traits and personalities (see also Hastie, 1989).

Thus, although most of our natural language concepts may have large clusters of characteristic features associated with them that yield stereotypes, prototypes, or other phenomenal, holistic, similarity spaces, in adults at least, most of these concepts also seem to "go beyond" the stereotype or the merely typical. (See also Armstrong, Gleitman, & Gleitman, 1983.) With many kinds, we tend to go beyond with theories that provide some explanation of why features causally interrelate. Feathers, wings, flight, and light weight do not just co-occur in birds, they all tend to mutually support the presence of each other and, by doing so, segregate the set of things known as birds into a natural kind; and our understanding of birds as a kind may require partial grasp of those causal relations that result in stable configurations such as birds. Murphy and Medin (1985) summarized this new perspective nicely as a need for "conceptual coherence."

It is these beliefs that allow us to make such powerful inductions about natural kinds given some set of properties, far more powerful ones than we are able to make given comparable properties with most nominal kinds and simple artifacts. Concepts for artifacts have some interesting commonalities with concepts for that subset of natural kinds known as biological kinds, as is shown later, but they also tend to differ from all natural kinds by not having elaborate theories overriding characteristic features as much as simple definitions or functional descriptions and less of an assumption of an essence. Some complex artifacts, such as computers and perhaps even televisions, start to blur this distinction and seem to become more essence possessing, but the generalization works quite well for simpler artifacts and most artifacts made prior to, say, 1600 A.D.

The emergence of this new consensus on the importance of intuitive theories in understanding concept structure, however, has created a new and perhaps much more profound controversy. Granting the need for such theories in describing the structure of mature concepts, there are deep and dramatic differences of opinion on how theory comes to constrain concept structure over the course of development. In this chapter, I explore some

different models of how the emergence of theoretical beliefs might come to constrain the acquisition and structure of concepts, as an attempt to describe one way in which structural constraints can guide cognitive development. I suggest that there is one view of how theories come to constrain concept structure and acquisition that seems to fit with a large body of traditional and current developmental data; I call it the "original sim." view. Then, however, I suggest that, for both principled and empirical reasons, the original sim. view may not be right and that a very different model is needed of how theories come to constrain concept growth, a view that has more general implications for understanding constraints on cognitive development.

Quine's Proposal

In a well-known essay on natural kinds, Quine (1977) offered a particularly clear statement of how theoretical beliefs can and cannot constrain concept acquisition. In essence, he argued that the young child starts life without any real theoretical beliefs but rather something much closer to an associative matrix laid over a set of sensory and perceptual primitives. Then, out of this primordial net of associations, theoretical beliefs emerge and come to restructure similarity and hence categories, concepts, and concept structure. The following quote nicely summarizes Quine's (1977) point of view:

> Between an innate similarity notion on spacing of qualities and a scientifically sophisticated one, there are all gradations. Science, after all, differs from common sense only in degree of methodological sophistication. Our experiences from early infancy are bound to have overlaid our innate spacing of qualities by modifying and supplementing our groupings habits little by little, inclining us more and more to an appreciation of theoretical kinds and similarities, long before we reach the point of studying science systematically as such. Moreover the latter phases do not wholly supersede the earlier; we retain different similarity standards, different systems of kinds, for use in different contexts. We all still say that a marsupial mouse is more like an ordinary mouse than a kangaroo, except when we are concerned with genetic matters. Something like our innate quality spaces continued to function alongside the more sophisticated regroupings that have been found by scientific experience to facilitate induction. (167–168) This development is a development away from the immediate, the subjective, animal sense of similarity to the remoter sense of similarity determined by scientific hypotheses and posits and constructs. Things are similar in the latter or theoretical sense to the degree that they are interchangeable parts of the cosmic machine revealed by science. (p. 171)

This makes clear why this is an "original sim." account of concept and theory development. Concepts are ultimately deeply dependent on theoretical beliefs for their internal structure and patterns of acquisition; but they

start out from an atheoretical original sim., or "animal sense of similarity" that is governed by domain-general mechanisms of learning.

Some Background Evidence for the Original Sim. View

Quine originally made his proposal more 20 years ago, and clear threads of his account can be seen to reach back much further in time. Thus, many empiricist philosophers, such as Locke and Hume, posited some sort of initial state like an original sim. and end states that are fully laden with knowledge and belief. There is also a long tradition of claims by psychologists that the young child organizes categories and concepts in a relatively atheoretical manner reflecting some general learning or abstraction procedure but then shifts to a more principled mode of concept organization. Vygotsky (1934/1986), for example, talked about an instance bound to principled shift for kinship concepts like "uncle," wherein a young child seemed to take all features that typically co-occurred with instances of uncle and more-or-less blindly tabulated them up to form an aggregate concept of uncle. The older child focused on a few principled relations governing bloodline relations and downplayed other highly characteristic features. Werner (1948) talked about a holistic-to-analytic shift that was somewhat similar to Vygotsky's, and even Piaget has, on occasion, made similar observations. Consider, for example, the following, albeit socioculturally dated, quote from Inhelder and Piaget (1964) on classification skills: ". . . making supper 'belongs with' a mother although it is hardly an essential property which she shares with all mothers. True, most mothers make supper; and we could think of these "belongings" as similarities. But such similarities are accidental rather than essential, since not all mothers make supper . . . The child is lumping a not quite essential attribute with the object its supposed to define" (pp. 36–37).

Across these and several other classic views there emerges a common theme of a dramatic qualitative shift in how concepts are structured. A shift from early representations in which all typically co-occurring properties and relations are tabulated to representations having a much tighter, more principled structure that is organized around a core set of interconnected beliefs, or what one might call theories. The early representations seem phenomenal and shallow and much like those organized by Quine's original similarity space.

Many of the older views tended to see this kind of shift as quite global and across the board, reflecting fundamental change in representational competency, perhaps a shift from young children being solely Roschean prototype abstractors (e.g., Rosch & Mervis, 1975) to older children also possessing more complex, but still domain-general, modes of learning and representation. This is still a popular view in some quarters, but it has been mostly overshadowed by a new emphasis on domain specificity.

Domain Specificity and Qualitative Shifts

A number of recent researchers, such as Chi, Hutchinson, and Robin (1989), Carey (1985), and Brown (1990), have argued that, although evidence for qualitative shifts across all domains is hard to come by, it seems much easier to make the case on a domain-by-domain basis. In my own work, I made this argument most extensively with respect to an apparent "characteristic to defining" shift (Keil, 1989; Keil & Batterman, 1984) in the acquisition of word meanings. This shift appears to be from representations based on holistic tabulations of all symptomatic or characteristic features to those where a few defining features predominate. Children are given descriptions in which either an instance has all the characteristic features of a concept but lacks critical defining features or, alternatively, an instance has the critical defining features but has highly atypical features as well.

For example, following Vygotsky, a person might be described who is 2 years old and makes a mess of Fred's toys and who is also the brother of Fred's father. The child is asked if said person is Fred's uncle. A second person might be described who has many of the most stereotypical features of uncles, such as a friendly disposition particularly to Fred, frequent appearance in Fred's household on holidays, the bearing of gifts for Fred, and reminiscing with Fred's father; but this person is also carefully described such that he could not possibly be related to Fred's parents. Younger children tend to deny the first person's status as an uncle and accept the second, whereas older children do just the opposite. This pattern also occurs for other terms, such as "island," "jail," "taxi," "lunch," and "advertisement"; but, contrary to older accounts, it does not occur at the same time for all concepts. Domain-specific shifts are quite common such that children shift to a reliance on defining features at a much earlier age for moral act terms than for cooking terms. Indeed, across a wide enough range of domains, characteristic-to-defining shifts seem to occur both in preschoolers and in novice to expert transitions in adults (see also Sera & Reittinger, 1990; Chi, Feltovich, & Glaser, 1981).

But although these characteristic-to-defining shift findings, as well as many of the older studies on conceptual change, provide strong support for domain specificity, they comprise only indirect evidence for an original-sim.-to-theory shift. They strongly suggest an original sim., but they are less able to indicate a shift to theory because they have carefully focused on special sorts of concepts that are not natural kinds: concepts for "nominal kinds" (see Locke, 1690, and Schwartz, 1977, for more on these contrasts between natural and nominal kinds). Such concepts have relatively clear definitions, which tend to be social constructs, and are just the sorts of concepts where the influence of theory on structure is likely to the least dramatic. We tend not to have elaborate theories of what makes something an island versus a peninsula. It is, as is commonly said, "simply a matter of definition."

Similarly, an uncle is usually designated by a small set of clear bloodline relations, not by a more elaborate theory. Thus, the normally intricate networks of beliefs that yield intuitive theories with natural kind concepts tend to be most impoverished with the more conventional concepts associated with nominal kinds.

Elsewhere Putnam (1975) has referred to these as "one-criterion terms," again contrasting them with natural kinds. There is arguably more theory impinging on these than might appear at first (see Lakoff, 1987a, 1987b, for example), but certainly the most compelling involvement of theory is with natural kinds. Consequently, the best evidence for shifts from original sim. to theory should arise out of assessments of children's concepts of natural kinds, for it is with natural kinds that theory is most influential. In the recent literature calling for attention to theory in describing concept structure, the predominant examples involve natural kinds, as was the case in Quine's original essay. Before exploring developmental work along these lines, however, a clearer analysis is needed of how concepts and theories might interrelate.

Theories and Concepts

Concepts of almost all sorts may "go beyond" the stereotype or the merely typical, but they may not do so in the same way for all kinds. For nominal kinds," such as uncles, triangles, and islands, the characteristic is transcended largely by a social/conventional construct that is quite close to what we think of classical definition consisting of necessary and sufficient features. (It is, in fact, rarely, if ever, actually easily decomposable to such features, but it certainly has that look.) We tend not to have rich sets of causal beliefs about why the core properties co-occur as they do. Moreover, we see little linkage between the core and the symptoms. That is, we do not see a highly structured essence that is intimately related to the characteristic features. The typical characteristic features of islands (e.g., palm trees, pirates' treasure) have little do to do causally with critical features of islands as, say, contrasted with peninsulas. Nominal kinds do not have a rich causal structure that is intrinsic to them and that is largely responsible for many of their typically associated properties. We don't have sciences of them and tend not to have rich theories invoking essences, because they are often to be considered either an arbitrary sort of convention or the product of human intentions (cf. Schwartz, 1977).

Essences are much more commonly associated with natural kinds; but, at the same time, when they are construed as necessary and sufficient features, they become problematic (e.g., Schwartz, 1977; Putnam, 1975). One therefore needs something else that is more fundamental than the merely typical but that is also not a simple definition. One possibility for what picks out natural kinds in the real world may be patterns of "causal homeostasis"

(Boyd, 1986). Roughly put, although most properties in the world may be ultimately connectable through an elaborate causal chain to almost all others, these causal links are not distributed in equal density among all properties. On the contrary, they tend to cluster in tight bundles separated by relatively empty spaces. What makes them cluster is a homeostatic mechanism wherein the presence of each of several features tends to support the presence of several others in the same cluster and not so much those in other clusters. Thus, the properties tend to mutually support each other in a highly interactive manner. To return to an example used previously, feathers, wings, flight, and light weight don't just co-occur; they all tend to mutually support the presence of each other, and, by doing so, segregate the set of things known as birds into a natural kind.

Boyd's claim is about natural kinds and what they are, not about psychology. At the psychological level, however, we may be especially sensitive to picking up many of these sorts of homeostatic causal clusters such that beliefs about those causal relations provide an especially powerful cognitive "glue," making features cohere and be easier to remember and induce on. This "adhesive" quality of beliefs about homeostatic relations may be roughly analogous to work on children's and adults' memories for stories showing more accurate and complete recall when episodes are causally connected rather than merely temporally connected (e.g., O'Brien & Myers, 1989; Stein & Glenn, 1979). Causal relations that provide a "story" unifying the frequently co-occurring elements of natural kinds may be powerful organizing components of concepts; and they may be especially powerful when the causal relations are structured homeostatically rather than in other causal ways, such as in linear chains. In addition, causal relations in general may be more effective than other noncausal but equally highly interconnected sets of relations such as the highly "systematic" clusters discussed by Gentner (1983) and Billman and Jeong (1989).

So, concepts for natural kinds may rely heavily on tightly connected sets of beliefs about the mechanisms responsible for the real world homeostasis that in fact partitions the world up into natural kinds. It is these beliefs that allow us to make such powerful inductions about natural kinds given some set of properties, perhaps more powerful ones than we are able to make given comparable properties with most nominal kinds and artifacts (Gelman, 1988). Of course, our concepts can hardly represent all such relations for most kinds as even our best theories often fail to do so; they may only need some set of interconnected causal beliefs even if they only partially, or perhaps even erroneously, describe the kind in question.

The different kinds can be construed as arrayed along a continuum from nominal kinds, such as islands and mortgages, to simple artifacts, such as tables and hammers, to complex artifacts, such as cars and computers, to natural kinds. Although there is no clear dividing line between one sort of

kind and another, it does seem clear that, as we move towards the natural kind end of the continuum, there is an increasing richness and internalization of causal homeostatic relations and a decreasing well-definedness as the cluster of causal relations itself becomes the essence rather than a simple one-line definition. There are richer causal/explanatory structures for the nominal kinds than there appear to be at first, but these tend to be more external to the kinds and are between those kinds and the social context in which they are embedded. For example, artifact properties do enter in a rich set of causal/explanatory relations with human intentions, culture, and ergonomic considerations, but these clusters do not seem as intrinsic to the kinds themselves.

Remember, again, that the psychological claim here is that part of our understanding of what natural kinds are is to have an appreciation of those causal/explanatory relations that help explain the mechanisms responsible for the emergence and maintenance of such clusters. The appreciation need not be complete, and probably never is, and may not even be accurate as illusory correlations illustrate. It mostly needs to be highly interconnected in such a way as to provide a mortar that cements individual properties and relations into a stable whole.

Constraints, the Original Sim., and Natural Kinds

The question at hand is whether theoretical constraints are emergent sorts of things that spring out of an associative matrix solely as a consequence of domain-general laws of learning, thus constraining concept development only to the extent that these theoretically driven similarities diverge from those of the original sim.. This account allows for dramatic qualitative shifts in manner of concept representation and acquisition that could occur on a domain-by-domain basis as theoretical relations are uncovered in each domain.

When I first started looking at how children come to understand natural kinds, it started to look as if the original sim. view might just be right and that it would fit the developmental tradition discussed previously. One technique that we have used extensively to assess children's natural kind concepts uses an operations paradigm where one changes all the salient characteristic features of one kind into those of another, contrasting kind. Thus, a raccoon might be turned into a skunk by dying and shaving its fur and teaching it to act like a skunk and hang around skunks and even putting inside it a sack of super-smelly yucky stuff to squirt out whenever it gets mad at other animals (or we change a tiger into a lion, a horse in to a zebra, a pine tree into an oak tree, or gold into lead). In these tasks, we changed all those features that would normally be mentioned by someone in a Roschean prototype task as prototypical of one kind into those that are prototypical of another kind. If these features are all there is to the concept, then the sort of thing described

should change as well; and sure enough, the younger children do say the animal is changed into a new kind. It seems as if they are simply organizing their concepts in terms of tabulations of typical features in a way predicted by the original sim. and that little else is involved.

The artifact pairs are judged changed by all ages. For example, when a bridge is transformed into something that looks like and functions as a table, almost everyone sees it to now be a table. It is important that these artifact transformations are caused by intentional agents who alter the features and functions with specific new goal states in mind. Transformations that are created more accidentally, or that involve new uses but with no feature changes, will naturally get more ambiguous responses, as the object's nature is more closely linked to the earlier intended function of its creator.

These sorts of studies, as well as others done with discoveries about properties, seem to support the notion of an atheoretic to theory-driven shift, with theory-based constraints on concepts and learning only emerging relatively late. Resistance to identity change by older children is interpreted as the emergence of a biological theory that overrides the characteristic feature cluster. It is also quite clear from these and other studies that the shift is not merely from what you can see to what you cannot see or changing response bias or some quirk of western culture. On the latter point, Jeyifous (1986) has shown that similar shifts occur among traditional nonliterate, nonwestern members of the Yoruba people in central Nigeria.

We have, therefore, developed a picture of concepts in which theoretical beliefs only come to influence concept structure as a result of a gradual accumulation of theoretical beliefs that are acquired through a domain general mechanism such as Quine's "trial and error learning," It seems almost trivially true that experts often learn the same body of information in different ways and with different endstate knowledge structures than novices; and there have been compelling demonstrations that such changes involve qualitative restructuring (e.g., Chi et al., 1981). The more provocative claim is that, contrary to novices who may have simpler or merely different theories, sufficiently young children have none at all, with the consequence that their representations are fundamentally different in kind from those of older children and adults. The remainder of this chapter asks if such a claim is warranted and what alternatives might be possible.

Some Theoretical Concerns About Original Sim and Qualitative Change

Potential problems with the original sim. account arise both at the theoretical and empirical levels. At the theoretical level, such a view requires that coherent theories be able to develop out of something like networks of associations, that interconnected sets of explanatory beliefs can rise out of

nothing more than probabilistic tabulations of features and relations. This notion falters when one recognizes that there are no persuasive accounts in any domain showing how this might occur. Perhaps the strongest claims along these lines are in the many studies summarized by Langely, Simon, Bradshaw, and Zytkow (1987).

Langely et al. assumed that important cases of theory acquisition can be described as a completely data-driven process of "Baconian induction," wherein theories are able to rise out of tabulations of raw data aided by sets of weak, domain-general heuristics (see also Holland, Holyoak, Nisbett, & Thagard, 1986). The controversial aspect to this research lies in the data that is supplied to the inductive device. The data is highly constrained in its input format and thereby may partly embody the theory supposedly being induced. Thus, Langely et al.'s computer simulations of theory discovery, $BACON_{1\text{-}6}$, all are provided highly structured patterns of input data. Langely et al. (1987) briefly address this problem by arguing that, historically, several important theories were constructed out of data that had been widely available for many years, such as with Mendeleev's discovery of the periodic table. Such instances are said to be "clear-cut cases of data-driven Baconian induction." (p. 24).

Something seems amiss in this conclusion about discoveries such as Mendeleev's, partly because no details are given of how they are achieved and partly because, if the event was purely data driven, it becomes even less clear why it took so long for someone to make the correct induction over that data. (Elsewhere, in a footnote (p. 23) Langely et al. seemed to acknowledge that the data as well must be "impregnated with theories," but this is dismissed as causing a hopeless chicken/egg problem.)

In addition to the problems of coming up with a noncontroversial successful simulation, there are reasons for suspecting that arguments for the necessity of a priori domain-specific constraints on the acquisition of simple concepts might apply in even more force for the acquisition of theories. Thus, even though Quine clearly wants no a priori constraints on the structure of theories, he simultaneously grants the clear need for innate constraints on initial categorizations to solve the problems of inducing the meanings of novel words. (See Keil, 1981, 1989, and Markman, 1989 for more extensive discussions of Quine's argument.) These innate "perceptual quality spaces" (Quine, 1960) are generally sympathetic to the empiricist tradition, in that Quine discussed them in terms of biases to see colors and shapes with the hope that more complex abstract conceptual relations will arise inductively out of these few sensory and perceptual primitives. Thus, ideally, there may be innate constraints on theories, but only those that follow from our sensory and perceptual biases, certainly not ones stated with reference to types of theories themselves.

The problem with this long-cherished empiricist account is explaining how such relations are "bootstrappable" up out of the sensory and perceptual qualia. In Quine's own writings, some of the posited innate quality spaces may demand far more central levels of cognitive bias than those granted by even the most charitable construals of sensory receptors. Evolution and natural selection have been frequently invoked by Quine as providing ways of narrowing down construal of such things as the physics of bodies and particles (Quine, 1981). Quine readily endorsed the view that we innately share with many animals a basic similarity space; but he may be off the mark in assuming that such a similarity space can be stated in purely sensory terms, even in many animals. Gallistel (1989) reviewed large bodies of evidence showing innate constraints on many animals' representations of space, time, and number, notions not easily reducible to sets of sensory primitives; and Spelke (1989, this volume) has shown how even the simplest ability to pick out objects, something Quine explicitly wants (Spelke, this volume) cannot be derivable from even the most sophisticated Gestalt principles. To the extent that animals such as rats and pigeons are governed by such "higher level" constraints, Quine's "animal sense of similarity" no longer advances an empiricist point of view.

Some of Quine's own examples seem to go beyond the sensory in profound ways. For example, in discussing Goodman's riddle of induction and problem of determining whether something is green or "grue" (green before the year 2000 and blue thereafter), Quine and Ullian (1973) stated that "our innate sensitivities have served us much better than purely random selection of traits would likely have served us, and that our animal faith bids us expect continuance of our good fortune" (p. 89). A preference for green as a "trait," and not grue, would not seem to be distinguishable at the level of the sensory receptors, for the receptors would serve equally well in picking out both types of traits. Deeper notions of what sorts of properties vary temporally and in what ways for what kinds are necessary.

Quine, therefore, must acknowledge the need for some constraints on concepts; and having granted such biases, it is not clear why they are not also needed for more complex concatenations of concepts, such as theories. Quine's original arguments for quality spaces were motivated largely by demonstrations of the innumerably large alternate sets of construals of the meanings of single words such as "rabbit," construals that would thwart any knowledge acquisition process without support from a constraining quality space. If such biases are needed to narrow down an otherwise hopeless search space of possible hypotheses about single word meanings, why should one not need even richer sets of biases on combinations of such meanings and on larger scale relational structures? The alternative is to argue that somehow the constraints at certain level of concepts of physical objects are adequate for

truncating the search space over all more complex structures; but no rationale is given for why the constraints should suddenly stop at the level of rabbits versus brief rabbit temporal slices.

More broadly, it is difficult to know how causal explanatory beliefs emerge out of the associative sorts of innate similarity spaces envisioned by so many empiricists. There are many different senses of cause that complicate these accounts (going all the way back to Aristotle's different kinds of causes), but problem remains largely independent of these different senses.

Empirical Challenges to the Doctrine of Original Sim.

There are also empirical reasons for wondering if younger children are really as atheoretical as they appear to be in these tasks. It is not clear, for example, that for even the youngest child, all salient features and correlations are equally noticed. If a young child has somehow seen robins only when his mother is around and sparrows when others are around as well, it is nonetheless highly unlikely that he will assume that such a mother alone / robin correlation is at all meaningful to the notion of robins. Perhaps there are theoretical biases from the start that discount the reasonableness of such correlations. One alternative to Quine suggested that children start off with one, or perhaps two, innate theories, out of which all others originate. One of clearest advocates of that alternative is Carey (1985).

In discussing some of the operations studies described earlier in this chapter Carey suggested that they may represent a theory-to-theory shift, in particular a shift from construing biological kinds in terms of a theory of behavior to construing them in terms of a theory of biology. Such a shift fits nicely with Carey's own work on changing patterns of inductions about biological kinds, wherein young children appear to project properties to other animals roughly on the extent to which they are behaviorally and psychologically similar to their most familiar animal, humans. Carey suggested that there are only two basic theories: one of behavior and one of the mechanics of physical objects of the sort that Spelke (this volume) has discussed. More broadly, we might call such views the "primal theory" views, the idea being that all later theories spring out of one or two seminal ones. A third alternative, discussed further hereafter, "the pluralistic theories view," advocates an initially more diverse set of biases that, from the start, guide the development of theories in several different domains. Both of these alternatives agree on one point. Children may not be so atheoretical as they seem in the tasks just described; perhaps they are not spineless phenomenalists helplessly bound to computing frequencies and correlations over sensory and perceptual primitives. If not, we need a different way of describing the changes in responses observed in the discoveries and operations studies as well as in more classical studies arguing for such things as holistic-to-analytic shifts.

Even if the youngest of children are constrained by content specific biases on theories, such constraints are not on their own adequate to model the apparent shifts described earlier. One also needs the additional assumption that concepts are always heterogeneous blends of theory and "associative" structure. The youngest child and the most sophisticated adult never relies solely on one to the exclusion of the other. As adults, even our most elaborate theories eventually run dry in their abilities to meaningfully distinguish one subkind from another, and even the youngest child may have some deeply held central theoretical beliefs that can override the characteristic for them as well. By this account, there may be no such thing as an original sim. except for completely artificial and meaningless concepts such as "blibs" being large, blue triangles with fuzzy textures.

With natural kinds, at least, there are theoretical biases from the start that constrain induction and learning. What develops in any domain may therefore not be the emergence of theory out of nothing but rather the continuous presence of theory that, as it changes or becomes superseded, reinterprets and adjusts similarity relations over more and more of the associative matrix on which it is overlaid. Concepts of kinds will seem to shift in qualitative ways, but perhaps only because they are reflecting the gradually increasing ability of a set of beliefs to explain correlational patterns that had previously only been knowable associatively. However, because we often presuppose common parts of the theory so strongly and completely with the young child, we may only see at first a shift from the associative residual to theory and thereby miss that there was an even more basic set of theoretical biases all along.

Going Beyond the Original Sim.

If younger children are able to go beyond the phenomenal by using properties that are closer to their core beliefs, one should be able to find cases where they override characteristic features just as much as older children and further show that they seem to do so on the basis of theoretical beliefs specific to the domain in question. A series of studies has been conducted with the goal of demonstrating such a process.

Intercategory Distance. One possible way to explore if more central properties can be used to override the merely typical is to present young children with transformations that cross not only species but also more fundamental boundaries, such as those between plants and animals and between biological kinds and artifacts. Even very young children may have beliefs about certain core properties of all animals and thus be unwilling to see an animal transformed into a nonanimal even if it is still a natural kind. To explore this possibility, children were given not only the familiar tigers-into-lions and horses-into-zebras cases, but they were also given examples of

porcupines being transformed into apparent cacti and toy mice into apparent real mice. In all cases, photographs of real exemplars were used, under the assumption that any tendencies to override the especially vivid characteristic features of the objects depicted by such photographs would be especially strong evidence for appeals to more principled knowledge.

When these transformations are described to kindergarteners, there is a strong divergence in response patterns as a function of whether the transformations cross fundamental category lines versus those of related species. The children consistently doubt that one can turn an animal into a nonanimal or vice versa while simultaneously maintaining that one can easily turn one animal into another animal. Thus, they may have more principled understandings of the differences between such categories as plants, animals, and other kinds but may not have as deep an understanding for the differences between such categories as species.

An alternative and sharply contrasting way of explaining such results is to assume that they simply represent a relatively atheoretical similarity metric, such that the more perceptually dissimilar two kinds are, the more resistant the child is to think there is change in kind after a transformation. These two views would make differing predictions about young children's judgements of similar versus dissimilar animals. With the atheoretical metric, there should be a gradual rise in judgement that kind is not changed as intercategory distance increases. For example, a child might happily judge a mouse to be transformable into a chipmunk, reluctantly accept a mouse transformation into a tarantula, but vehemently deny that it could be transformed into a mouse-like pile of brown moss. Alternatively, with more principled beliefs about animals as a kind in general, one might expect to see little or no rise as animal/animal similarity decreased (e.g., from mouse chipmunk to mouse tarantula); but then a dramatic rise as animals are changed into other sorts of kinds, such as plants and rocks.

In a follow-up study, we have conducted just such a comparison, and the results are clear. Species that are not at all similar by adult metrics, such as spiders and mice, and butterflies and fish, are just as intertransformable as more similar pairs, such as zebras and horses (Keil, 1989). Judgments that transformations do not change the kind are only seen when transformations cross the animal/other boundaries, thereby favoring the idea that more principled beliefs, perhaps about biological kinds, are helping these younger children override the characteristic.

Varying the Type of Transformation The intercategory distance studies therefore reveal at least one way in which a child may look like he or she is a phenomenalist, because we have presupposed perhaps the most central theoretical distinction and only tested a more peripheral and later developing

one. But perhaps there are more subtle analyses that will provide further detail of what develops. Rather than manipulating distance between category pairs for the same sorts of transformations, one might instead manipulate the kinds of transformations used on precisely the same sets of pairs, under the assumption that different sorts of transformations will be more or less legitimate ways of changing kinds. Thus, some ways of transforming a horse into a zebra-looking and zebra-behaving thing might involve mechanisms that are clearly outside of anybody's realm of biologically relevant changes, whereas others might seem species-changing to all but the most educated adult. We manipulated transformations of animal properties in the following three ways:

1. A change that did not alter existing features but merely covered them up with costumes, masks, etc.

2. A change that did alter features in exactly the same way as in the previous studies except that the change was indicated as potentially temporary if one failed to periodically repeat an abbreviated version of the transformation.

3. A change that altered the same features and behaviors, but did so in ways that might be biologically plausible to even some adults. Some sort of internal intervention, such as a pill or injection, occurred early in the animal's life such that it gradually grew up to look and act like the other kind in the queried pair.

The idea here is to be able to reproduce the shift at any point of development by adjusting the nature of transformation such that it gets at the heart of the child's current theory. If theory and associative structure are thereby shown to be intermingled at several different ages, the notion of a distinct original sim. stage followed by a theory stage becomes less tenable.

In such studies, even preschoolers seem to overrule an original sim. when their core beliefs are accessed (the costume case), and even fourth graders are indecisive when one changes things more central to their beliefs (the injection case). Remember also that, in all cases, exactly the same photographs are used, such that the only differences are the kind of mechanism underlying the transformation. It therefore seems that one can recapitulate the shift at virtually any point in development that one wishes.

In addition, this study suggests, for the first time, that the younger children might not simply be construing all biological things in psychological or behavioral terms; perhaps they are different from older children because they have weaker theories specific to biological kinds. The kindergartner who rejects the costume as changing kind of animal doesn't seem to be relying on

any psychological or behavioral differences as much as some beliefs about what sorts of mechanisms are legitimately responsible for the manifestation of typical properties and what relations govern the individuation of biological kinds. It is difficult to advance a set of behavioral differences that could drive their strong intuitions that the horse in a zebra costume is still a horse.

Choosing Between the Original Sim. and Its Two Alternatives

The empirical studies described so far do not conclusively confirm one account of how concepts and theories become intertwined They do make some suggestions, however, and point to some possible ways in which such issues might be resolved. One cannot yet rule out a much earlier occurring original sim. that simply gives way to theory much sooner than was traditionally thought; but it is now possible to see how many apparent cases of original sim. are not bona fide. Moreover, I have suggested that it is also difficult to come up with in-principle accounts of how theory emerges out of associative networks. Rather than being the favored default option, the original sim. view may now be that which carries the burden of proof.

Choosing between the primal and pluralistic alternatives is more complicated, but the last study on mechanisms of transformation and property manifestation does raise doubts about accounts in which all early concepts of biology are couched only in behavioral terms. To further explore the pluralistic alternative, a series of studies is needed that systematically compares behavioral/psychological forms of explanation against other types, such as the purely mechanical and the specifically biological. Studies are now approaching the question from several directions.

One set of studies recently completed (Vera & Keil, 1988) suggests that preschoolers are capable of making inductions about biology on a conceptual base other than a naive psychology. Carey's pioneering work on preschoolers' inductions about biological kinds suggests that, when asked if various animals, for example, eat, the children assented to the extent that the animals were behaviorally similar to known exemplars that ate, especially humans. This is a robust and easily replicable finding, but it may not mean that there is only one belief system available for such inductions. When preschoolers are asked to make the same inductions except that the features are embedded in contexts that suggest the relevance of biological relations, they show dramatically different patterns of induction that indicate an appreciation of animals as natural kind.

Many more studies of this sort are needed to fully understand how specific theories emerge. With biology, we are currently conducting studies on beliefs about disease, digestion, and inheritance of properties in attempt to see if there are invariant biases that guide belief formation about such kinds, biases

that aren't reducible to those that spawn beliefs about psychological kinds. More broadly, this strategy might be repeated in any theoretical domain that seems to be culturally and historically universal.

Conclusions

The studies I have just described are obviously a long way from telling us what, if any, the original theories are. Younger children, cross-cultural studies, and other converging measures are needed before we can make any strong claims about what is present from the start. But these studies do show us how one can mistake growth of a theory in one domain for a qualitative shift from an original sim. If concepts are always a blend of a kind of associative matrix overlaid by causal beliefs, and if much of concept development consists of a theory's increasing interpretation of that matrix, then we may often witness an apparent shift from an original sim., because we presuppose and assume the relevant causal beliefs so automatically that we only focus on the remaining associative structure that becomes infiltrated with belief. It is only by systematically considering the full range of causal relations that are implicit in many natural kind concepts that we begin to see theory-driven adjustments of the similarity space in even the youngest of children.

In summary, the studies in this chapter point towards the following four themes:

1. There is never a pure original sim. for any natural domain . . . if one finds it at all, it is in the realm of totally artificial and meaningless synthetic concepts such as a "glub" being a blue, small, triangle with a fuzzy texture.

2. There is never pure theory, either. Even the most sophisticated adult theories of natural phenomena will run dry, and we have fall-back mechanisms for then representing what's left over. Thus, we can store vast amounts of correlational information to use as a base for further development. Sometimes, in all our fuss to talk about concepts embedded in theories, it seems as if all information must be couched in a particular theory or else we cannot store it all. This cannot be right.

3. Intuitive theories constrain concept acquisition in two ways. Initially, they constrain by virtue of whatever biases we have to prefer some classes of mechanisms over others; but they also constrain as they become further elaborated and take on values that may be quite idiosyncratic to local cultures and belief systems. Obviously, there must be complex and important interactions between these initial biases and the patterns of data they encounter. Concept structure is neither completely data driven nor completely theory driven. This fits well with the more general view of constraints argued for in this volume.

4. Even though a child may have never entertained a single thought about a mechanism underlying some phenomena, we shouldn't be led to conclude that he or she doesn't have very strong preferences to prefer highly specific classes of explanations or clusters of causal beliefs over others, illustrating the central question of how these are represented and what it means to "have a theory" versus a set of pre-theoretic biases.

Work on concepts, intuitive theories, and their constraining relations in development are importantly related to broader concerns about structural constraints on cognitive development. Domain specificity figures powerfully in discussions of how theories come to restructure concepts, and it is equally important in discussions of other sorts of constraints. Whether the content is number, syntax, naive physics, or birdsong, it is essential to understand the extent to which constraints are tailored to particular kinds of knowledge or are more general guidelines on all kinds of knowledge. I have suggested that domain-specific constraints are of critical importance in understanding how theories come to influence concepts, and similar emphases are seen in several other chapters in this book.

A second important theme is the difference between constraints as skeletal frameworks versus fully articulated restrictions. As has been repeatedly stressed, theories do not exert a fully deterministic influence on concept structure. There is clearly some diversity in concept structure, even when similar broad-based theories apply. If, for example, there are invariant biases on a naive physics, these biases can hardly be used to precisely predict the full set of beliefs about physical objects. They provide a framework that makes some relations more cognitively natural than others; but this framework may influence such judgments of naturalness in a generative fashion so that indefinitely large sets of beliefs are biased rather than one explicit set. The analogies to other putative constraints, such as those on syntactic rules, are evident.

A final theme stresses the intrinsically interactional nature of constraints. I have argued that there may well be predetermined biases on theory construction in not just one or two domains but possibly other domains as well. At the same time, it is clear that these biases are best understood as governing interactions between the child and its environment. They do not state that some set of beliefs are unknowable in all contexts or that others must invariably appear; rather, they suggest that, across certain ranges of "normal" environments, such biases are evident. Thinking of constraints in this manner defuses heated and needless controversies over the appropriateness of terms such as *innate* and *learned* while at the same time recognizing that we are biological organisms that may well have evolved adaptations for building knowledge representations about sets of regularities in our physical, social, and formal worlds.

ACKNOWLEDGMENTS

Preparation of this paper and some of the research reported on herein was supported by NIH grant 1-R01-HD23922. Much thanks to Susan Carey and Rochel Gelman for comments on an earlier draft of this manuscript.

REFERENCES

Armstrong, S., Gleitman, L., & Gleitman, H. (1983). What some concepts might not be. *Cognition, 13*, 263–308.

Asch, S. (1952). *Social psychology*. New York: Prentice-Hall.

Billman, D., & Jeong, A. (1989, November). Systematic correlations facilitate learning component rules in spontaneous category formation. Paper presented at the 30th meeting of the Psychonomic Society, Atlanta, GA.

Boyd, R. (1986). *Natural kinds, homeostasis, and the limits of essentialism*. Unpublished paper presented at Cornell University, Ithaca, NY.

Brown, A. (1990). Domain-specific principles affect learning and transfer in children. *Cognitive Science, 14*, 107–134.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psycho-diagnostic signs. *Journal of Abnormal Psychology, 74*, 272–280.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representations of physics problems by experts and novices. *Cognitive Science, 5*, 121–152.

Chi, M. T. H., Hutchinson, J. E., & Robin, A. F. (1989). How inferences about novel domain-related concepts can be constrained by structured knowledge. *Merrill Palmer Quarterly, 35*, 26–62.

Gallistel, C. R. (1989). Animal cognition: The representation space, time, and number. *Annual Review of Psychology, 40*, 155–189.

Gelman, S. A. (1988). The development of induction within natural kind and artifact categories. *Cognitive Psychology, 20*, 65–90.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*, 2.

Hastie, R. (1989, November). *Complex social impressions*. Paper presented at the 30th annual meeting of the Psychonomic Society, Atlanta, GA.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT Press.

Inhelder, B., & Piaget, J. (1964). *The early growth of logic in the child*. New York: Norton.

Jeyifous, S. (1986). *Antimodemo: Semantic and conceptual development among the Yoruba*. Unpublished dissertation, Cornell University, Ithaca, NY.

Keil, F. C. (1981). Constraints on knowledge and cognitive development. *Psychological Review, 88*, (3), 197–227.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Bradford Books.

Keil, F. C., & Batterman, N. (1984). A characteristic-to-defining shift in the development of word meaning. *Journal of Verbal Learning and Verbal Behavior, 23*, 221–236.

Lakoff, G. (1987a). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.

Lakoff, G. (1987b). Cognitive models and prototype theory. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 63–100). Cambridge: Cambridge University Press.

Langely, P., Simon, H. A., Bradshaw, G. L., & Zytkow, J. M. (1987). *Scientific discovery.* Cambridge, MA: MIT Press.

Locke, J. (1975). *An essay concerning human understanding* (P. H. Nidditch, Ed.). Oxford: Clarendon Press. (Original work published 1690)

Markman, E. (1989). *Categories and word meaning in children.* Cambridge, MA: MIT Press.

Medin, D. L., & Shoben, E. J. (1988). Context and structure in conceptual combinations. *Cognitive Psychology, 20,* 158–190.

Murphy, G. L., & Medin, D. (1985). The role of theories in conceptual coherence. *Psychological Reviews, 92,* 289–316.

O'Brien, E. J., & Myers, J. L. (1989). The role of causal connections in the retrieval of text. *Memory and Cognition, 15,* 419–427.

Putnam, H. (1975). The meaning of meaning, In H. Putnam (Ed.), *Mind, language and reality* (Vol. 2, pp. 215–271). London: Cambridge University Press.

Quine, W. V. O. (1960). *Word and object.* Cambridge, MA: MIT Press.

Quine, W. V. O. (1977). Natural kinds. In S. P. Schwartz (Ed.), *Naming, necessity, and natural kinds* (pp. 155–175). Ithaca, NY: Cornell University Press.

Quine, W. V. O. (1981). Five milestones of Empiricism. In W. V. O. Quine (Ed.), *Theories and things* (pp. 89–108). Cambridge, MA: Harvard University Press.

Quine, W. V. O., & Ullian, J. S. (1973). *The web of belief.* New York: Random House.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7,* 573–605.

Sera, M. D. & Reittinger, E. L. (1990). Developing definitions of objects and events in English and Spanish speakers. Unpublished manuscript, University of Minnesota.

Schwartz, S. P. (1977). *Naming, necessity and natural kinds.* Ithaca, NY: Cornell University Press.

Spelke, E. S. (1989). The origins of physical knowledge. In L. Weiskrantz (Ed.), *Though without language.* Oxford University Press.

Stein, N., & Glenn, C. (1979). An analysis of story comprehension in elementary school children. In R. Freedle (Ed.), *New directions in discourse processing* (Vol. 2, pp. 186–211). Norwood, NJ: Ablex.

Vera, A., & Keil, F. C. (1988). *The development of induction about biological kinds: The nature of the conceptual base.* Paper presented at the 1988 meeting of the Psychonomic Society, Chicago, IL.

Vygotsky, L. S. (1986). *Thought and language.* (E. Hantmann and G. Vakar, Trans.) Cambridge: MIT Press. (Original work published 1934)

Werner, H. (1948). *Comparative psychology of mental development* (2nd ed.). New York: International Universities Press.