The Theory theory of Concepts

Daniel A. Weiskopf

The Theory theory of concepts is a view of how concepts are structured, acquired, and deployed. Concepts, as they will be understood here, are mental representations that are implicated in many of our higher thought processes, including various forms of reasoning and inference, categorization, planning and decision making, and constructing and testing explanations. The view states that concepts are organized within and around theories, that acquiring a concept involves learning such a theory, and that deploying a concept in a cognitive task involves theoretical reasoning, especially of a causal-explanatory sort.

The term "Theory theory" derives from Adam Morton (1980), who proposed that our everyday understanding of human psychology constitutes a kind of theory by which we try to predict and explain behavior in terms of its causation by beliefs, intentions, emotions, traits of character, and so on. The idea that psychological knowledge and understanding might be explained as theory possession also derives from Premack & Woodruff's famous 1978 article, "Does the chimpanzee have a theory of mind?" A "Theory theory" in general is thus a proposal to explain a certain psychological capacity in terms of a (tacit or explicit) internally represented theory of a domain. The Theory theory of concepts, however, goes beyond the mere claim that we possess such theories, saying in addition that some or all of our concepts are constituted by their essential connections with these theories.

The origins of the Theory theory involve several converging lines of investigation. First, it arose as part of a general critique of the formerly dominant prototype theory of concepts; second, it was an empirically-motivated response to the shortcomings of the developmental theories of Piaget and Vygotsky; and third, it involved applying ideas from Kuhn's philosophy of science to explain phenomena having to do with the development of cognition in individuals. While the theory has often been vaguely formulated, due in large part to the open-endedness inherent in the central notion of a theory, there are substantial bodies of empirical evidence that underlie the main tenets of the view. In particular, the Theory theory has been responsible for largely displacing the notion that cognitive development starts from a simple base of perceptual primitives grouped together by similarity. Rather, it is guided by domain-specific explanatory expectations at many stages, and these expectations can be seen to function in adult reasoning and categorization as well. While strong versions of the Theory theory have been subject to numerous objections, these contributions endure and continue to shape the best existing models of higher cognition.

Table of Contents

1. Background

The Theory theory emerged in part as a reaction to existing trends in the psychology of concepts and categorization, which during the late 1970's was dominated by the prototype theory of concepts. Exemplar models were also being developed during this time, but the prototype theory encapsulates many of the views which were the foils against which the Theory theory developed its main assumptions.

Prototype theory derives in large part from the work of Eleanor Rosch and her collaborators (Rosch, 1977; Rosch & Mervis, 1975; see Smith & Medin, 1981 for historical perspective and Hampton, 1995 for a canonical statement of the view). These theories assume that concepts represent statistical information about the categories that they pick out. The concept TREE represents the properties that people take to be typical of trees: they have bark, they can grow to be relatively tall, they have green leaves that may change color, they have a certain silhouette, birds often nest in them, they grow potentially edible fruits, and so on. These comprise the tree prototype (or stereotype).

This stereotype is acquired by a process of abstraction from examples: individual trees are perceptually salient parts of the environment, and by repeated perception of such category instances, one gradually forms a summary representation that ‘averages’ the qualities of the trees one has observed. This summary is often represented as a list of features that belong to category members. Properties that are more frequently perceived in the instances will be assigned a greater feature weight in the prototype. This process of concept acquisition is often portrayed as a passive one.

Finally, novel objects are categorized as falling under a prototype concept in virtue of their similarity to the prototype—that is, by how many features they share with it (weighted by those features’ importance). Similarity computations also explain other phenomena, such as the fact that some objects are better examples of a category than others (flamingos and penguins are atypical birds since they lack most of the prototypical BIRD features).

The prototype theory has several characteristics which made it a fitting target for Theory theorists. First, it suggests that concepts have a basically superficial nature. Often, though not invariably, features in prototypes were assumed to be readily perceivable. Prototype theory was thus affiliated with a certain empiricist bent. This was reinforced by the fact that prototypes are acquired by a simple statistical-associative process akin to that assumed by classical empiricists. Second, prototype theory involved a relatively impoverished account of conceptual development and deployment. Concepts passively adjust themselves to new stimuli, and these stimuli activate stored concepts in virtue of their resemblances, but there is little role for active revision or reflective deployment of these concepts. In the wake of the anti-empiricist backlash that gave rise to contemporary cognitive science, particularly in cognitive-developmental psychology, these assumptions were ripe for questioning.

2. The Theory theory

a. Origins of the view

The Theory theory itself has a somewhat complicated origin story, with roots in a number of philosophical and psychological doctrines. One is the reaction against stage theories of cognitive development, particularly Piagetian and Vygotskian theories. Stage theories propose that children's cognitive development follows a rigid and universal script, with a fixed order of transitions from one qualitatively distinct form of thought to another taking place across all domains on the same schedule. Each stage is characterized by a distinctive set of representations and processes. In Piaget's theory, children move through sensorimotor, preoperational, concrete operational, and formal operational stages from birth to roughly ages 11 or 12 years old. Similarly, Vygotsky held that children move from a stage of representing categories in terms of sensory images of individual objects, through a stage of creating representations of objectively unified categories, and finally a stage of categories arranged around abstract, logical relationships.

While Piaget and Vygotsky's stage theories differ, both hold that early childhood thought is characterized by representation of categories in terms of their perceivable properties and the inability to reason abstractly (causally or logically) about these categories. Early childhood cognition, in short, involves being perceptually bound. While the empirical basis and explanatory structure of these theories had been challenged before (see R. Gelman & Baillargeon, 1983 and Wellman & Gelman, 1988 for review), Theory theorists such as Carey (1985), Gopnik (1988, 1996), Gopnik & Meltzoff (1997), and Keil (1989) went beyond providing disconfirming evidence and began to lay out an alternative positive vision of how cognitive development proceeds.

A second root of the Theory theory derives from philosophy of science, particularly from Kuhn's account of theory change and scientific revolutions. Kuhn's view is too complex to summarize here, but two aspects of it have been particularly influential in developmental psychology. One is Kuhn's notion that theory change in science involves periods of 'normal science', during which a mature theory is applied successfully to a range of phenomena, and periods of 'paradigm shifting'. These paradigm shifts occur when counterevidence to a theory has built up beyond a certain threshold and it can no longer be adequately modified in response, consistent with its not becoming intolerably *ad hoc*. In paradigm shifts, new explanatory notions and models take center stage, and old ones may be pushed to the margins or adopt new roles. New practices and styles of experimentation become central. These changes are relatively discontinuous compared with the gradual accumulation of changes and modifications characteristic of normal science.

Second, connected with this notion of a paradigm shift is the Kuhnian doctrine of incommensurability. This is the idea that when new theories are constructed, the central explanatory concepts of the old theory often change their meaning, so that a claim made before and after a paradigm shift, even if it uses the same words, may not express the same proposition, since those words now express different concepts. The concept of mass as it existed in pre-relativistic physics no longer means the same thing—indeed, we now need to distinguish between uses of 'mass' that pick out rest mass and those that pick out relativistic mass. Often this involves creating new concepts that cannot be captured in the conceptual vocabulary of the old theory, differentiating two concepts that were previously conflated, or coalescing two previously distinct concepts into one. In all of these cases, the expressive vocabulary of the new conceptual scheme is not equivalent to that of the old scheme. Theory theorists have often adopted both the Kuhnian claim about paradigm shifts as a model for understanding certain phenomena in development, and the associated claim of semantic or conceptual incommensurability (Carey, 1991).

A third root involves what Keil (1989) dubs the rejection of ‘Original Sim’. Original Sim is roughly the view of category structure and learning suggested by prototype theory, particular of an empiricist variety. This view is most clearly defended by Quine (1977), who proposes that children begin life with an innate similarity space that is governed by perceptual information, and over time begin to develop theoretical structures that supersede these initial groupings. On this empiricist perspective, children’s first concepts should be bundles of perceptual features, typically consisting of intrinsic rather than relational properties, and categorization should be simply a matter of matching the perceived features of a novel object to those of the concept. Inductive inferences concerning a category are within reach so long as they are confined to these observable properties, and objects share inductive potential to the extent that they are similar to the perceptual prototype. Moreover, these perceptual prototypes are assumed to be acquired by statistical tabulation of observed co-occurrences in the world, in a relatively theory-free way; seeing that certain furry quadrupeds meow is sufficient for constructing a CAT concept that encodes these properties. It is only at later stages of development that concepts reflect understanding of the hidden structure of categories, and come to enable inductions that go beyond such similarities.

Fourth, Theory theory is often motivated by the hypothesis that certain concepts (or categories) have a kind of *coherence* that makes them seem especially non-arbitrary (Murphy &

Medin, 1985; Medin & Wattenmaker, 1987). The categories of diamonds, sports cars, or otters seem to be relatively 'coherent' in the sense that their members bear interesting and potentially explainable relations to one another: diamonds are made of carbon atoms whose organization explains their observable properties, otters share a common ancestry and genetic-developmental trajectory that explain their phenotype and behavior, etc. On the other hand, the category of things on the left side of my desk, or things within 100 feet of the Eiffel Tower, or things that are either electrons or clown wigs, do not. They are simply arbitrary collections.

Feature-based theories of concepts, such as prototype theory, seem to have particular difficulty explaining the phenomenon of coherence, since they are inherently unconstrained and allow any set of properties to be lumped together to form a category, whereas our concepts often appear to represent categories as involving more than merely sets of *ad hoc* co-instantiated properties. They include relations among these properties, as well as explanatory connections of various sorts. We don't merely think of sports cars as expensive artifacts with four wheels, big engines, a sleek shape, and bright coloration, which make a loud noise as they roar past at high speed. These features are explanatorily connected in various ways: their shape and engine size contribute to their speed, their engines explain their noisiness, their speed and attractive appearance explain their costliness, and so on. Insofar as these explanatory relations among properties are represented, concepts themselves more coherent, reflecting our implicit belief in the worldly coherence of their categories. Theories are the 'conceptual glue' that makes many of our everyday and scientific concepts coherent, and models of concepts that fail to accord theories an important role are missing an account of a crucial phenomenon (however, see Margolis, 1999 for detailed criticism of this notion).

From this survey, it should be clear that the development of Theory theories of concepts has been driven by a host of different motivations and pressures. Hence there exist many flavors of the view, each with its own distinctive formulation, concerns, and central phenomena. However, despite the fact that the view lacks a canonical statement, it possesses a set of family resemblances that make it an interesting source of predictions and a robust framework for empirical research, as well as a unified target of criticism.

b. Theories defined

The first essential posit of these views is the notion of a mentally represented theory. Theories are bodies of information (or, as psychologists and linguists sometimes say, bodies of knowledge) about a particular domain. Such theories have been posited to explain numerous psychological capacities: linguistic competence results from a theory of the grammar of English or Urdu; mental state attribution results from a theory of mind; even visual perception results from a theory of how 3-D objects in space behave in relation to the observer. But theories are not just any body of information held in memory. What makes theories distinctive or special? Keil (1989, p. 279) called this “the single most important problem for future research” in the Theory theory tradition.

Gopnik & Meltzoff (1997, pp. 32-41) give what is probably the most comprehensive set of conditions on theories. These conditions fall into three categories: *structural*, *functional*, and *dynamic*. Structurally, theories are abstract, coherent, causally organized, and ontologically committed bodies of information. They are abstract in that they posit entities and laws using a vocabulary that differs from the vocabulary used to state the evidence that supports them. They are coherent in that there are systematic relations between the entities posited by the theory and the evidence. Theories are causal insofar as the structure that they posit in the world to explain observable regularities is ordinarily a causal one. Finally, they are ontologically committed if the entities that they posit correspond to real kinds, and also support counterfactuals about how things would be under various non-actual circumstances. Some of these conditions are also advanced by Keil (1989, p. 280), who proposes that causal relations are central to theories, especially where they are homeostatic and hierarchically organized.

Functionally, theories must make predictions, interpret evidence in new ways, and provide explanations of phenomena in their domain. The predictions of theories go beyond simple generalizations of the evidence, and include ranges of phenomena that the theory was not initially developed to cover. Theories interpret evidence by providing new descriptions that influence what is seen as relevant or salient and what is not. And crucially, theories provide explanations of phenomena, understood as an abstract, coherent causal account of how the phenomena are produced and sustained. Theories are essentially related to the phenomena that make up their domain; hence in Keil's developed view, there is a key role for associative relations in providing the raw data for theoretical development as well as a 'fallback' for when theories run out (Keil, 1989, p. 281).

Last, theories are not static representations, but have dynamic properties. This follows from the fact that they develop in response to, and may gain in credibility or be defeated by, the empirical evidence. The sorts of dynamic properties that characterize theories include: an initial period involving the accumulation of evidence via processes of experimentation and observation, the discovery of counterevidence, the possible discounting of such evidence as noise, the generation of ad hoc hypotheses to amend a theory, the production of a new theory when an old one has accumulated too much contrary evidence or too many ugly and complicated auxiliary amendments.

c. Concepts in theories versus concepts as theories

Once the central explanatory construct of a mental theory is clear, two varieties of the Theory theory need to be distinguished. These views differ on the nature of the relationship between concepts and theories.

On the *concepts in theories* view, concepts are the constituents of theories. Theories are understood as something like bodies of beliefs or other propositional representations, and these beliefs have concepts as their constituents. The belief that electrons are negatively charged is part of our theory of electrons, and that belief contains the concept ELECTRON as a part (as well as HAS NEGATIVE CHARGE). The set of ELECTRON-involving beliefs that meet the sorts of constraints laid out in section 2b constitute one's theory of electrons. These beliefs describe the sorts of things electrons are, how they can be expected to behave, how they are detected, how they relate to other fundamental physical entities, how they can be exploited for practical purposes, and so on. For the concepts in theories view, concepts function much like theoretical terms.

On the *concepts as theories* view, on the other hand, the constituency relations run the opposite direction. Concepts themselves are identified with miniature theories of a particular domain. For instance, Keil (1989, p. 281) proposes that "[m]ost concepts are partial theories themselves in that they embody explanations of the relations between their constituents, of their origins, and of their relations to other clusters of features." So the concept ELECTRON would itself be made up of various theoretical postulates concerning electrons, their relationship to other particles, their causal propensities which explain phenomena in various domains of physics, and so on. Concepts are not terms *in* theories, they are *themselves* theories.

As stated, the concepts in theories view is scarcely controversial. If people possess mentally represented theories at all, then those theories are composed of beliefs and concepts, and so at least some of our concepts are embedded in theory-like knowledge structures. Call this the *weak concepts in theories* view. A *strong concepts in theories* view, on the other hand, says that not only are concepts embedded in theories, but they are also *individuated* by those theories. Carey (1985, p. 198) seems to hold this view: “Concepts must be identified by the roles they play in theories.” This is just to say that what makes them the very concepts that they are is their relationships (inferential, associative, causal, explanatory, etc.) with the other concepts and beliefs in the theory. There are many ways of carving out different notions of inferential or theoretically significant roles for concepts to play, but on all of them, concepts are constituted by their relations to other concepts and to the evidence that governs their conditions of application.

A consequence seems to be that if those relationships change, or if the theory itself changes in certain respects, then the concepts change as well. The change from a view on which atoms are the smallest, indivisible elements of matter to one on which atoms are made up of more fundamental particles might represent a sufficiently central and important change that the concept ATOM itself is no longer the same after such a transition takes place; similarly, perhaps the victory of anti-vitalism entailed a change in the concept LIFE from being essentially linked with a particular irreducible vital force to being decoupled from such commitments. Notice that this consequence also applies to the concepts as theories view. If a concept is identified with a theory (rather than being merely embedded in it), it seems as if, *prima facie*, any change to the theory is a change to the concept.

The concepts as theories view poses separate difficulties of its own. On this view, concepts are extremely complex data structures composed of some sort of theoretical principles, laws, generalizations, explanatory connections, and so on. What status do these have? A natural suggestion is to regard all of these as being beliefs. But this view is straightforwardly incompatible with a view on which concepts are the constituents of beliefs and other higher thoughts. It is mereologically impossible both for concepts both be identified with terms in theories and with theories themselves. We would need some other way of talking about the representations that make up beliefs if we choose to regard concepts as simply being miniature theories.

Despite the differences between these two views, the empirical evidence taken to support the Theory theory does not generally discriminate between them, nor have psychologists always been careful to mark these distinctions. As with many debates over representational posits, the views in question generate differing predictions only in combination with supplementary assumptions about cognitive processing and resources. However, there may be theoretical reasons for preferring one view over the other; these will be discussed further in section 5.

3. Support for the Theory theory

a. Cognitive development

Much of the support for the Theory theory comes from developmental studies. Carey (1985) largely initiated this line of research with her investigations of children’s concepts of ANIMAL, LIVING THING, and kindred biological notions. For example, she found that major changes occur in children’s knowledge of bodies and their functioning from age 4 years to age 11. The youngest children understand eating, breathing, digesting, etc., mainly as human behaviors, and they explain them in terms of human needs, desires, plans, and conventions. Over time, children build various new accounts of bodies, initially treating them as simple containers and finally differentiating them into separate organs that have their own biological functions. In Carey's terms, young children start out seeing behavior as governed by an intuitive psychological theory, out of which an intuitive biology develops (1985, p. 69).

The centrality of humans to young children's understanding of living things can be seen in several studies. Four- and 5-year olds are reluctant to attribute animal properties—even eating and breathing—to living beings other than humans. When asked to name things that have various properties of living things, children overwhelmingly pick 'people' first, followed by mammals, and then a few other scattered types of creatures. The primacy of people in biology carries over to judgments of similarity, with adults displaying a smooth gradient of similarity between people and other living things and 6-year olds seeing a sharp dividing line between people and the rest of the animal kingdom, including mammals. Finally, in inductive projection tasks people are clearly paradigmatic for 4-year olds: if told that a person has an organ called a spleen, they will project having a spleen to dogs and bees, but rarely the opposite. By age 10, people are no longer unique in this respect. So young children's theory of life is focused initially around humans as the paradigm exemplars, and only later becomes generalized as they discover commonalities among all animals and other living things. Indeed, the very concept LIVING THING comes to be acquired as this knowledge develops.

Keil (1989) added to the evidence with many striking results concerning how children's concepts of natural kinds, nominal kinds, and artifacts develop from kindergarten onwards. He finds compelling evidence for what he initially called a 'characteristic-to-defining' shift in conceptual structure. Characteristic features are akin to prototypes: compilations of statistically significant but possibly superficial properties found in categories. Defining features, on the other hand, are those that genuinely make something the kind of thing that it is, regardless of how well it corresponds to the observed characteristics.

In a series of *discovery studies*, Keil (1989) gave children descriptions of objects that have the characteristic features belonging to a natural kind, but which were later discovered to have the (plausible) defining features of a different kind; e.g., an animal that looks and acts like a horse but which is discovered to have the inside parts of cows as well as cow parents and cow babies. While at age 5, children thought these things were horses, by age 7 they were more likely to think them cows, and adults were nearly certain these were cows. Defining features based on biology (internal structure, parentage) come to dominate characteristic features (appearance, behavior).

In a related series of *transformation studies*, children heard about a member of a natural kind which underwent some sort of artificial alterations to its appearance, behavior, and insides; e.g., a raccoon that was dyed to look like a skunk and operated on so that it produces a foul, skunk-like odor. Five-year olds thought these transformations changed the raccoon into a skunk, while 7- year olds were more resistant, and 9-year olds were nearly sure that such changes in kind weren’t possible. This effect was notably stronger for biological kinds than mineral kinds; however, children at all ages strongly resisted the idea that a member of a biological kind could be turned into something from a different ontological category (e.g., an animal cannot be turned into a plant). Finally, some kinds of transformations are more likely to change a thing’s kind: among 5-year olds, alterations to internal or developmental features along with permanent surface parts are more effective than temporary surface changes or costumes, and internal changes retain their influence until at least age 9.

In Keil's view, this shows that children may start out with a comparatively impoverished theory of what makes something a member of a biological kind (or a mineral kind, social kind, or artifact kind), but this theory is enriched and deepened with more causal principles governing origins, growth, internal structure, reproduction, nutrition, and behavior. As this network of causal principles becomes more enriched they recognize that the category members are defined by the presence of these theoretically significant linkages rather than by the more superficial features that initially guided them. 'Primal theories' develop into more mature folk theories in different domains according to their own time course.

Finally, Gopnik & Meltzoff (1997) survey a range of domains to argue for the early emergence of theories. To take one example, they argue that children's understanding of objects and object appearances starts off as highly theoretical and develops in response to new experience until they achieve adult form. Six-month olds, for instance, fail to search for objects that are hidden behind screens, and they show no surprise when an object moves behind a screen, fails to appear at a gap in the middle of the screen, but then appears whole from behind the other side of the screen. These behaviors only emerge at 9 months. Gopnik & Meltzoff explain this change by claiming that the infants come to understand that occlusion makes objects invisible. Until 12 months, however, they continue to make the 'A-not-B' error, which involves searching for an object under the first occluder it disappeared behind, rather than the last one. They ascribe this failure to children's adherence to an auxiliary hypothesis of the form: objects will be where they appeared before. This rule is abandoned when it comes to conflict with the evidence and the child's developing theory of object behavior. In addition, from ages 12 to 18 months, children begin to systematically play ('experiment') with hiding and invisible displacement, suggesting that they are interested in generating evidence about this developing cognitive domain. This in turn strengthens the analogy between cognitive development and active theorizing by adult scientists.

b. Adult categorization, inference, and learning

Murphy & Medin (1985) argued in largely abstract fashion that categorization should be seen as a process of explaining why an exemplar belongs with the rest of a category. A man who jumps into a swimming pool while fully clothed at a party is plausibly drunk, even though these are not features of drunks in general—or they certainly are not stored as such in one's default DRUNK concept. Theories and explanatory knowledge are required to focus on the relevant features of categories in a variety of tasks and contexts. Research with adults has tended to support this perspective.

One significant piece of evidence that comports with the general Theory theory perspective is the causal status effect (Ahn & Kim, 2000). The effect is the tendency of participants to privilege causally deeper or more central properties in a range of tasks including categorization and similarity judgment. For example, if people are taught about a person who has a cough caused by a certain kind of virus, and then given two other descriptions, one which matches in the cause (same virus) but not the effect (runny nose), and another that matches in the effect (cough) but not the cause (different virus), common causal features make exemplars more similar. Matching causal features can even override other shared features in categorization. If taught about an example with a cause that produces two effect features and two other examples, one of which shares the cause only and the other of which shares both effects, a majority of participants group the common cause exemplar with the original, even though they differ in most features.

Murphy (2002) reviews an extensive body of evidence showing that background knowledge has a pervasive effect on category learning, categorization, and induction. To take two examples, consider artificial category learning and category construction. In learning studies, participants are given two categories that are distinguished by different lists of features. The features that describe a more 'coherent' category in which the features are very plausibly related to each other (e.g., 'Made in Norway, heavily insulated, white, drives on glaciers, has treads') were played against those that describe a more 'neutral' category. Participants found the coherent categories much easier to learn, and retained more information about them. Similarly, if given the ability to freely sort these items into categories they tended to group the coherent category members together even when they shared only a single feature. Background knowledge concerning the likely relationships among these features plays an essential role in learning and categorizing, even when it is not explicitly brought up in the experiment itself. This further undermines the prototype theory's account of learning as a process of atheoretical tabulation of correlations.

4. Relations to other views

a. Relations to essentialism

The Theory theory is closely related to psychological essentialism (henceforth just 'essentialism'), the claim that people tend to represent categories as if they possessed hidden, non-obvious properties that make them the sorts of things that they are and that causally produce or constrain their observable properties (Medin & Ortony, 1989). These essences need not be actually known, but may be believed to exist even in the absence of detailed information about them. Concepts may include either conjectures as to what their essential properties might be, or else blank ‘essence placeholders’ that govern in the absence of these as-yet-unknown essential properties. Commitment to essences may be viewed as a kind of theoretical commitment, insofar as essences are causally potent but unobserved properties that structure and explain observable properties of categories. More generally, it is the commitment to there being a certain kind of causal structure underlying the categories we commonly represent.

There is a large body of evidence that supports the psychological essentialist hypothesis (Gelman, 2003, 2004; see Strevens, 2000 for criticism). For example, children’s inductions are governed by more than superficial resemblances among objects. In a standard inductive projection paradigm, participants are presented with a triad of pictures of objects only two of which perceptually resemble each other (e.g., a leaf, a leaf-shaped insect, and a small black insect) and two of which share a verbal label (e.g., both insects, while dissimilar, are called ‘bugs’, and the leaf is called a ‘leaf’). They are then told that one object of the resembling pair has a certain property and asked to project the property to the third object. By 30 months, children will project properties on the basis of labeled category membership rather than similarity. This effect does not depend on the precise repetition of the verbal label (i.e., synonyms work just as well), and it tends to be more powerful in natural biological kinds than in artifacts. Even among 16- to 21-month olds one can find similar effects: behaviors displayed with one sort of toy animal (barking, chewing on a bone, etc.) will be imitated with a perceptually dissimilar animal if they are given a matching label. This suggests that induction is not entirely governed by superficial properties even among very young children.

Children may entertain more specific hypotheses about what the underlying category essences are as well. In Keil’s transformation studies, some participants, when debriefed, maintained that parentage was important to determining kind membership. In a number of studies, Gelman and her collaborators (see, e.g., Gelman & Wellman, 1991) have shown that among 4- to 5-year olds, insides have a special theoretical role to play. Children can distinguish similar-appearing objects (pigs and piggy banks) from those that have similar insides, and they judge that removing a creature's insides both removes its category-typical behaviors and also makes it no longer the same kind of thing. Removing outsides or changing a transitory property has little effect on membership or function.

These studies provide further evidence that the Original Sim has at best a weak grip on young children. Moreover, they reinforce the claim that categorization can sometimes be dominated by an early-emerging understanding of biology that treats stereotypical properties as non-dispositive. Gelman's own robust psychological essentialism includes further claims such as that category boundaries are invariably taken to be sharp rather than fuzzy, and that essences invariably focus on purely internal properties. Whatever the status of these additional claims, the broader moral of the essentialism literature is in line with the proposals made by Theory theorists. Children come prepared to learn about deeper causal relations in many domains and they readily treat these relations as important in categorizing and making inductions.

b. Relations to causal modeling approaches

In recent years much attention has focused on the role of causality in cognition, and consequently theories of cognitive performance that emphasize causal modeling have gained prominence. The idea that concepts might be identified (at least in part) with causal models has grown out of this tradition.

The theory of causal models is a formally well-developed and quantitatively precise way of describing probabilistic and causal dependency information, particularly in graphical form (for accessible introductions, see Gopnik & Shulz, 2007; Glymour, 2001; Sloman, 2005). Briefly, a causal model of a category depicts part of the relevant causal information about how things in the domain are produced, organized, and function. A causal model of a bird notes that it has wings, a body, and feathers, but also encodes the fact that those features causally contribute to its being able to fly; a causal model of a car depicts the fact that it is drivable in virtue of having wheels and an engine, that it can transport people because it is drivable, and that it makes noise because of its engine. These structures can be represented as sets of features connected by arrows, which indicate when the presence of one property causes or sustains (and therefore makes more probable) the presence of another. These directed causal graphs provide one possible representational format for concepts.

For example, Chaigneau, Barsalou, & Sloman (2004) have proposed the HIPE theory of artifact categorization, which states that artifacts are grouped according to their Historical role, the Intentions of the agents that use them, their Physical structure, and the Events in which they participate. On HIPE, artifact concepts are miniature causal models of the relations among these properties, all of which may potentially contribute to making something the kind of artifact that it is. Similar sorts of models could be developed for natural kind concepts. Indeed, essentialism itself is one form that a causal model can take: the essence is the ‘core’ of the concept, and it causally produces the more superficial features. Causal model theory is a generalization of this idea that allows these graphs to take many different forms.

Causal model theory is best seen as one form that the Theory theory can take (Gopnik & Schulz, 2004; Rehder, 2003). It shares that view’s commitment to causal-explanatory structure being central to concepts. While it is tied to a more specific hypothesis about representation than Theory theory in general (the formalism of directed causal graphs), this is also a strength, since these models are part of a well-developed framework for learning and processing. Causal model theory gives the Theory theory the resources to develop more wide-ranging and detailed empirical predictions concerning categorization, induction, and naming.

It is also worth noting that causal model theory may give the concepts as theories view the resources to answer the mereological objection it faces. The components of causal models can be seen as features representing properties, connected by links representing causal relations. Many models of concepts take them to be complex structures composed of features in this way. If we see causal models as miniature theories, then we can view concepts as theories if we identify them with such models. Adopting this approach eliminates any potential problems about concepts being both the constituents of beliefs and also being composed of beliefs.

5. Objections to the Theory theory

a. Holism

The holism objection focuses on the fact that the individuation conditions for concepts are closely tied to those for theories. They are holistic, meaning that a concept's identity depends on its relations to a large set of other representational states. This position is suggested by Murphy & Medin's comment that "[i]n order to characterize knowledge about and use of a concept, we must include all of the relations involving that concept and the other concepts that depend on it" (1985, p. 297). This gives rise to problems concerning the stability of concepts. The objection may be put as follows. Suppose concepts are identified by their relation to theories. Then changes in theories entail changes in concepts: if $C_1,\ldots,C_n$ are constituted by their relation to T_1, and T_1 changes into T_2, then at least some of $C_1,\ldots,C_n$ will have to change as well, so long as the changes in the theories occurs in the parts that contribute to individuating those concepts. And it is part of the developmental and dynamical account of the Theory theory that such transitions in theories take place. So according to the Theory theory, concepts are unstable; they change over time, so that one does not have the same concepts before a revision in theory that one has afterwards.

The conclusion is particularly objectionable if one assumes that there will be many changes to theories, so that concepts also change frequently. But there are reasons to want concepts to be more stable than this. First, one wants to be able to compare concepts across individuals with different theories. A young child may not have the fully developed LIFE concept, but she and I can still have many common beliefs about particular living things and their behavior, even if she does not represent them as being alive in the way that I do (that is, even if her understanding of life is impoverished relative to mine). Second, the rationality of theory change itself depends on some intertheoretic stability of concepts. Rejecting theory T_1 may involve coming to believe that belief B formulated using T_1's concepts is false. So now that I believe T_2 I reject B. But if changing from T_1 to T_2 involves changing the concepts involved in B, I can no longer even formulate that belief, since I now lack the required conceptual resources. So we are at a loss to describe the rational nature of the transition between theories.

What this suggests is that belief attributions are often stable across theory changes; or at least, not every change in one's background theory should change many or all of one's concepts (and hence beliefs). Some sort of doxastic insulation is required. The problem is that concepts are individuated by their roles, which in turn are determined by the causal, inferential, and evidential roles of the propositions that contain them, and these are precisely what change as theories do (Fodor, 1994; Margolis, 1995).

This problem faces both the strong concepts in theories view and the concepts as theories view, but the weak concepts in theories view is immune to it, since it allows that concepts may participate in theories without being individuated by them. Two responses to the holism objection are typical. First, some Theory theorists (e.g., Gopnik & Meltzoff, 1997) have embraced it. It is, they suggest, not implausible that young children are to a certain degree incomprehensible to adults, as would be predicted if their world view is incommensurable with ours (Carey, 2009, p. 378). Second, others have attempted to avoid this conclusion by distinguishing respects in which concepts may change (such as narrow content or internal conceptual role) and respects in which they may remain stable (such as wide content or reference). This dual-factor approach is also adopted by Carey (2009). The unstable respects are those that differ with background theories, while the stable respects provide continuity so that concepts can be reidentified across changes and differences in view. The success of this approach depends on whether the stable respects can do the relevant explanatory work needed in psychological explanation and communication.

b. Compositionality

A representational system is compositional if the properties of complex symbols are completely determined by the properties of the simpler symbols that make them up, plus the properties of their mode of combination. So predicate logic is compositional, since the semantic value of 'Fa' is determined by the semantic values of the predicate 'F' and the individual constant 'a'. Similarly 'Fa & Fb' is semantically determined as a function of 'Fa', 'Fb', and the interpretation of conjunction. Many have argued that thought is compositional as well (Fodor, 1998), which entails that the properties of complex concepts derive wholly from the properties of their constituents.

If thought is compositional, and concepts are the constituents of thoughts, then whatever concepts are must also be compositional. So if concepts are (or are individuated by) theories, then theories must similarly be compositional. However, there are good reasons to think that theories are not compositional. A standard example is PET FISH. The concept FISH might come from the theory of folk biology, while PET might derive from a theory of human social behavior (since keeping pets is a sociocultural fact about humans). If the strong concepts in theories view is right, their content is determined by their inferential role in each of these theories. But PET FISH has a novel inferential role that is not obviously predictable from those roles taken individually. Instances of PET FISH, for example, are typically thought to live in bowls and feed on flakes, neither of which is true of pets or fish in general. This information is not derived from one's 'pet theory' or 'fish theory'. It is therefore not compositional. The same point can be made about the concepts as theories view. If one's causal models of pets and fish do not somehow encode this information in the features that make them up, then it cannot be derived compositionally by putting them together. Since examples like this can be multiplied indefinitely, the Theory theory cannot account for the general compositionality of thought.

While many psychologists have simply ignored these concerns, several responses are possible. Here are two. First, one can divide concepts into two components, a stable compositional element and a non-compositional element (Rips, 1995). The compositional element might be thought of just as a simple label, while the non-compositional element includes theoretical and prototypical information. One part has the job of accounting for concept combination, the other has the job of accounting for categorization and inductive inference.

Second, one can try to weaken the compositionality requirement. Perhaps concepts are required only to be compositional in principle, not in practice; or else compositionality might be viewed as a fallback strategy to be employed when there is no other information available about the extension of a complex concept (Prinz, 2002; Robbins, 2002). Whichever approach one takes, the compositionality objection highlights the fact that while the Theory theory has impressive resources for explaining facts about development and concept deployment, concept combination is more challenging to account for.

c. Scope

The scope objection is one that faces nearly every theory of concepts. In general, where such a theory proposes an identification of the form 'concepts are K', where 'K' is a kind of mental structure or capacity, the question can be raised: are *all* concepts like this? Or are there cases where someone might possess the relevant concept but not possess K? For instance, if concepts are prototypes, then there must be the right sort of prototype for every concept we can use in thought. A theory has satisfactory scope if there exists the right sort of K for every concept that we are capable of entertaining.

For the Theory theory, the problem seems to be that there are *too few* theories. We have concepts such as CAR, COMPUTER, GIN, LEMUR, and NIGHTSTICK. Perhaps for some of these we have theories, at least of a highly sketchy nature. But it is less clear that we have these for other concepts. One might have the concept HIGGS BOSON (in virtue of reading newspaper articles about the Large Hadron Collider) but have essentially no interesting knowledge of the Standard Model of particle physics. One might have the concept TRUE but not have a theory of truth. One might have the concept BILLIARDS but know nothing of the game's rules or conventions ('that thing they play in the UK that resembles pool'). If the Theory theory identifies each concept with a domain-specific theory, these scope challenges are serious. Denying that we have these concepts in virtue of lacking the relevant knowledge is unappealing.

One possible response is to restrict the scope of the Theory theory itself. Carey (1985) takes this tack. She does not think that every concept must be associated with a proprietary theory. Rather, concepts are embedded in relatively large scale theories of whole cognitive domains: "there are only a relatively few conceptual structures that embody deep explanatory notions—on the order of a dozen or so in the case of educated nonscientists. These conceptual structures correspond to domains that might be the disciplines in a university: psychology, mechanics, a theory of matter, economics, religion, government, biology, history, etc." (Carey, 1985, p. 201). This approach, favored by other domain theorists, gives this version of the concepts in theories view a slight advantage over the concepts as theories view, since the latter is more clearly threatened by the scope objection. A defender of the concepts as theories view might fall back to the position that even very sketchy or minimal understanding of the causal principles at work in a category can count as a theory, so even in these cases we meet the minimal concept possession conditions, since our understanding is often precisely this superficial (Rozenblit & Keil, 2002).

d. Disanalogies between development and science

The Theory theory relies heavily on the notion that what children do in constructing their knowledge of the world is quite literally like what scientists do in producing, testing, and revising the theories that constitute scientific knowledge. This implies that there is substantial cognitive continuity across development, so that infants and young children, along with older children and adults, employ the same theory-construction mechanisms that operate on prior theoretical representations plus new evidence to produce revised and, with luck, improved theories.

Many have challenged this picture on the grounds that what children do is not in fact sufficiently similar to what scientists do for them to be seen as instances of the same cognitive or epistemic process. These complaints are summarized by Faucher, Mallon, Nazer, Nichols, Ruby, Stich, & Weinberg (2002). They argue that scientific theory revision is a process that is inseparable from a host of cultural factors. For example, there are norms governing how one ought to gather and weigh evidence, as well as how one should revise one’s beliefs, and these govern the practice of science differently across times and cultures. Moreover, theories are usually socially transmitted (in the classroom, the lab, and in less formal contexts) along with these norms. So in science, society and mind interpenetrate in such a way that individual cognition needs to be receptive to external sources of authority, both with respect to theoretical knowledge and epistemic norms. The simple picture of theory revision as involving only initial theories and evidence should be rejected.

There are at least two possible responses to this anti-individualistic argument. One is to argue that while these social factors play a role in adult science, the essential core of scientific practice remains the adjustment of theories under the influence of evidence. Normative factors can eventually come to help us perform these tasks better or in ways that fit in more productively with the surrounding culture, but the basic mechanism of evidence-based revision must be present in any case. And the evidence suggests that it is operative even before these social factors have an effect. A second response would be to argue that this picture is in fact a correct and welcome revision to the overly simplistic view originally proposed by Theory theorists. We should expect there to be substantial cultural influences on children's cognition, and some of the cross-cultural studies cited by Faucher et al. provide evidence in favor of this hypothesis. So we should enrich the Theory theory view of children's early cognition, not abandon it entirely.

6. Conclusion

The Theory theory consists of many interrelated claims about concept individuation, structure, development, and processing. The claim that development of concepts and domain knowledge in children is driven by causal-explanatory expectations, perhaps of an essentialist sort, has been most extensively investigated. While there are some attempts to explain these data by appeal to empiricist principles (Smith, Jones, & Landau, 1996), the Theory theory has strong support here. Studies with adults also suggest that causal information is often important to categorization. The behavior of both adults and children has been characterized using the framework of causal models, enabling Theory theorists to frame their view in a formally precise way. Many of the assumptions that trouble the account, such as the strong concepts in theories view that generates the problems of holism and incommensurability, turn out not to be essential to its empirical success. The greatest problem the view faces may be one of scope, but this challenge is arguably faced by all other theories of concepts currently in contention (Machery, 2009; Weiskopf, 2009). Whether or not a thoroughgoing Theory theory perspective is ultimately vindicated, its key insights will have to be incorporated by any future comprehensive theory of concepts.

7. References

Ahn, W., & Kim, N. S. (2000). The causal status effect in categorization: An overview. In D. L. Medin (Ed.), *Psychology of Learning and Motivation*, Vol. 40 (pp. 23-65). New York: Academic Press.

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

Carey, S. (1991). Knowledge acquisition: enrichment or conceptual change? In S. Carey & R. Gelman (Eds.), *The Epigenesis of Mind* (pp. 257-291). Hillsdale, NJ: Erlbaum.

Carey, S. (2009). *The Origin of Concepts*. Oxford: Oxford University Press.

Chaigneau, S.E., Barsalou, L.W., & Sloman, S. (2004). Assessing the causal structure of function. *Journal of Experimental Psychology: General, 133,* 601-625.

Faucher, L., Mallon, R., Nazer, D., Nichols, S., Ruby, A., Stich, S., & Weinberg, J. (2002). The baby in the labcoat: Why child development is an inadequate model for understanding the development of science. In P. Carruthers, S. Stich & M. Siegal (Eds.), *The Cognitive Basis of Science* (pp. 335-362). Cambridge: Cambridge University Press.

Fodor, J. (1994). Concepts: A potboiler. *Cognition, 50,* 95-113.

Fodor, J. (1998). *Concepts*. Oxford: Oxford University Press.

Gelman, R., & Baillargeon, R. (1983). A review of some Piagetian concepts. In J. H. Flavell and E. Markman (Eds.), *Cognitive Development: Vol. 3* (pp. 167-230). New York: Wiley.

Gelman, S. (2003). *The Essential Child*. Oxford: Oxford University Press.

Gelman, S. (2004). Psychological essentialism in children. *Trends in Cognitive Sciences, 8,* 404-409.

Gelman, S., & Wellman, H. (1991). Insides and essences: Early understandings of the nonobvious. *Cognition, 38,* 213-244.

Glymour, C. (2001). *The Mind's Arrows*. Cambridge: MIT Press.

Gopnik, A. (1988). Conceptual and semantic development as theory change. *Mind and Language, 3*, 197-217.

Gopnik, A. (1996). The scientist as child. *Philosophy of Science, 63*, 485-514.

Gopnik, A., & Meltzoff, A. (1997). *Words, Thoughts, and Theories*. Cambridge: MIT Press.

Gopnik, A., & Schulz, L. (2004). Mechanisms of theory-formation in young children. *Trends in Cognitive Science, 8*, 371-377.

Gopnik, A., & Schulz, L. (Eds.) (2007). *Causal Learning*. Oxford: Oxford University Press.

Hampton, J. A. (1995). Similarity-based categorization: The development of prototype theory. *Psychologica Belgica, 35*, 103-125.

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

Machery, E. (2009). *Doing Without Concepts*. Oxford: Oxford University Press.

Margolis, E. (1995). What is conceptual glue? *Minds and Machines, 9*, 241-255.

Margolis, E. (1999). The significance of the theory analogy in the psychological study of concepts. *Mind and Language, 10*, 45-71.

Medin, D., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and Analogical Reasoning* (pp. 179-195). Cambridge: Cambridge University Press.

Morton, A. (1980). *Frames of Mind*. Oxford: Oxford University Press.

Murphy, G. (2002). *The Big Book of Concepts*. Cambridge: MIT Press.

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

Medin, D., & Wattenmaker. (1987). Category cohesiveness, theories, and cognitive archeology. In U. Neisser (Ed.), *Concepts and Conceptual Development* (pp. 25-63). Cambridge: Cambridge University Press.

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

Prinz, J. (2002). *Furnishing the Mind*. Cambridge: MIT Press.

Quine, W. V. (1977). Natural kinds. In S. Schwartz (Ed.), *Naming, Necessity, and Natural Kinds* (pp. 155-175). Ithaca: Cornell University Press.

Redher, B. (2003). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 1141-59.

Rips, L. (1995). The current status of research on concept combination. *Mind and Language, 10,* 72-104.

Robbins, P. (2002). How to blunt the sword of compositionality. *Nous, 36*, 313-334.

Rosch, E. (1977). Human categorization. In N. Warren (Ed.), *Advances in Cross-Cultural Psychology: Vol. 1* (pp. 1-49). London: Academic Press.

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

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 26,* 521-562.

Sloman, S. (2005). *Causal Models*. Oxford: Oxford University Press.

Smith, E. E., & Medin, D. (1981). *Concepts and Categories*. Cambridge: Harvard University Press.

Smith, L., Jones, S., & Landau, B. (1996). Naming in young children: A dumb attentional mechanism? *Cognition, 60*, 143-171.

Strevens, M. (2000). The essentialist aspect of native theories. *Cognition, 74*, 149-175.

Weiskopf, D. (2009). The plurality of concepts. *Synthese, 169*, 145-173.

Wellman, H., & Gelman, S. (1988). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology, 43*, 337-375.