

Categorisation, causation, and the limits of understanding

Frank C. Keil

Yale University, New Haven, CT, USA

Although recent work has emphasised the importance of naïve theories to categorisation, there has been little work examining the grain of analysis at which causal information normally influences categorisation. That level of analysis may often go unappreciated because of an “illusion of explanatory depth”, in which people think they mentally represent causal explanatory relations in far more detail than they really do. Naïve theories therefore might seem to be irrelevant to categorisation, or perhaps they only involve noting the presence of unknown essences. I argue instead that adults and children alike effectively track high-level causal patterns, often outside awareness, and that this ability is essential to categorisation. Three examples of such pattern-tracking are described. The shallowness of our explanatory understandings may be further supported by a reliance on the division of cognitive labour that occurs in all cultures, a reliance that arises from well-developed abilities to cluster knowledge in the minds of others.

Categorisation is influenced by a wide range of types of information creating a diversity and richness that is often underestimated and which poses challenges for how we are to understand our abilities to have effective mental representations of categories. We have long recognised the importance of frequency and correlational information in influencing our formation and use of categories, and there is a temptation to reduce all of categorisation to such patterns through appeals to parsimony. This perspective, however, has been challenged in more recent times by those who argue that “theory-based” information also plays a critical role above and beyond that played by tabulations of frequencies and correlations.

Requests for reprints should be addressed to Frank Keil, Department of Psychology, Yale University, P.O. Box 208205, 2 Hillhouse Avenue, New Haven, CT 06520-8205, USA. E-mail: frank.keil@yale.edu.

Preparation of this paper and some of the research described therein was supported by National Institutes of Health Grant R01-HD23922 to Frank Keil. Many thanks to Paul Bloom, James Hampton, and three anonymous reviewers for comments on earlier versions of this paper and to Lisa Webb for help in manuscript editing.

© 2003 Psychology Press Ltd

<http://www.tandf.co.uk/journals/pp/01690965.html> DOI: 10.1080/01690960344000062

I will consider how causal information supports theory-based effects. The complexity and variety of causal patterns, in turn, will make it clear that people cannot possibly track all causal patterns associated with categories. Moreover, people grossly overestimate their own and others' abilities to know causal relations. This "illusion of explanatory depth" might be taken as an indication of the empty and ineffectual nature of intuitive theories. I will argue, however, that there is a different, often more implicit way in which people track causal structure, that powerfully influences concepts and categorisation.

CAUSATION, COMPLEXITY, AND CATEGORISATION

The process of learning about most categories critically involves noticing how often various properties occur and co-occur. Frequencies of instances of categories and of properties can have a powerful influence on later judgements of category membership, both in terms of the speed of such judgements and in terms of one's confidence about category membership (Hampton, 2001; Smith & Medin, 1981). In some cases, however, equally frequent properties seem to be weighed quite differently from each other, and in others, equal correlations are treated differently. One way of understanding such effects is to assume that our perception of the importance of features and correlations is also influenced by our understanding of how and why features and features clusters occur as they do (Murphy & Medin, 1985). This perspective has become popular in the cognitive science literature and has been called the "theory theory" (Gopnik & Wellman, 1994).

In the theory theory, concepts and the categorisation behaviour arising from them are both influenced by theories about how features are related. For example, blackness is probably a more frequent property of tyres than is roundness (given the many flat and partially deflated tyres in the world), yet roundness seems much more central to the concept of a tyre. This notion of theoretical centrality has been invoked to explain a wide variety of experimental results concerning concepts and categories (Murphy & Medin, 1985; Lin & Murphy, 1997; Murphy & Allopenna, 1994; Rehder & Hastie, 2001; Wisniewski & Medin, 1994; Ahn & Kalish, 2000; Sloman, Love, & Ahn, 1998; Murphy, 2002). Theory influences seem to be further supported by arguments that patterns of conceptual change across the development and the history of a science must be understood in terms of the theories in which concepts are embedded (Kuhn, 1977; Keil, 1989; Carey, 1991; Barrett, Abdi, Murphy, & Gallagher, 1993).

Although disagreements remain about the extent to which theory-like or explanatory relations do influence membership decisions about categories

and the speed with which one makes such decisions, most acknowledge some degree of influence. The challenge lies in describing the nature and extent of the relevant theoretical or background knowledge. For this discussion, I focus primarily on knowledge of causal relations. It is certainly possible to have theories that discuss no causal relations (e.g., in mathematics, or in some mathematical models of physical phenomena); but much of science, and most folk science, centers on causal relations.

There are at least three ways in which concepts might be related to theories: concepts are theories themselves, concepts are parts of theories, or concepts are influenced by theories (Prinz, 2002). Viewing concepts as theories has the attractive feature of explaining how conceptual change and theory-change are linked. Seeing concepts as parts of theories saves one from difficult issues concerning the compositionality of concepts. Finally, having concepts influenced by theories is the most conservative claim, capturing an influence effect that is implicit in the other two views but making no additional commitments. The arguments of this paper require no more than this third view.

If causal relations influence how we weight a property, how deeply must we master those causal relations? My concept of dogs might include the idea that they have hair to insulate them against cold. But that explanation, in turn, depends on an explanation of why hair was the means for insulation as opposed to blubber, or another way of insulation. That issue, in turn, could lead to questions about the nature of insulating materials, how they work, and how they might be produced in the most efficient fashion by a biological organism. Other questions might arise as to whether mammals are subject to different constraints concerning insulation needs and methods for providing insulation when compared with other animals. Still further questions occur about the distinctive natures of mammals, of ecological niches and thermodynamics. Full theoretical understanding creates a problem of causal explanatory holism, in which almost all the natural sciences in all their details seem to be required to fully explain any particular explanatory belief.

This problem of an ever-expanding chain of supporting explanations is clearly surmounted in both folk and laboratory science. Virtually no practicing scientists claim to understand all the causal chains behind the phenomena that they study. They adopt certain ideas, such as, that hair is good insulator, and then use those ideas in further reasoning without requiring a deeper understanding of how hair insulates. Moreover, this practice appears to be effective, given that the sciences do improve in their abilities to make predictions. In the biological and cognitive sciences, and in much of engineering, we seem to decompose a phenomenon or system into functional units and then analyse how those units interact to create the phenomenon in question. This functional-analytical approach (Lycan,

2002) succeeds because, in most cases, the functional units do not need to be further decomposed for an explainer to gain genuine insight.

The level of incompleteness of intuitive theories may also vary considerably across individuals. There are surely large individual differences in terms of how well people know the causal pathways underlying phenomena such as diseases, kinds of living things, or complex devices, with a corresponding potential for substantial disagreements over the relevant members of categories. If all of one's theoretical knowledge relevant to an entity influenced one's concept-learning and use, there should be more variation among individuals in categorisation than is normally observed. Thus, a nuance in a person's understanding of causal factors for a disease might cause him to reject a set of symptoms as an instance of a disease while another person would accept them as a member of that disease category.

The theory theory therefore offers a way of choosing among equally high frequencies and of deciding which correlations are causally meaningful (e.g., Murphy & Medin, 1985). However, its ability to simplify categorisation appears to be compromised when there is no clear way to decide how much and what type of causal understanding is needed. This problem is one reason why Fodor, in his criticism of cognitive science's approach to concepts, is so dubious of the concepts-as-theories view (Fodor, 1998). Theory-based influences are potentially unbounded, raising questions about how theories could be an effective structure for concepts if all causal relations are brought to bear. In addition, people vary considerably in the depth of their causal understandings, yet that variation does not map neatly onto individual differences in category use (see also Prinz, 2002).

The existing literature rarely considers the details of the theories that influence concept-acquisition and use and categorisation; and, when examples are provided, their lack of detail is illuminating. One frequently employed example suggests that features of birds such as wings, hollow bones, and feathers are mentally linked because they are understood as converging to support flight (e.g., Murphy & Medin, 1985). Similarly, curvedness is perceived as more important to boomerangs than to bananas because it is thought to be more causally central to explaining the nature of boomerangs (Medin & Shoben, 1988). In all these cases, however, the real level of the causal analysis is often surprisingly shallow, in comparison to what one thinks it is. Shape is important to boomerangs because it is considered causally critical to explaining their unique patterns of flight. This shape centrality seems to be a widespread "theoretical belief" (Medin and Shoben, 1988). Yet, very few people could explain why the bent shape of a boomerang actually makes it more likely to return than a straight one (see Walker, 1979, for a fuller explanation). People do not really have a

complete theory of boomerang flight; instead, they have a conviction that shape will be central to any such account. Similarly, in knowing that wings enable the flight of birds, people use very simple ideas of wings “holding the bird up” without really understanding how the wing shape achieves lift (Murphy, 2002). In most cases, people use simple causal schemas to guide their judgements about categories, and these schemas seem quite distant from theories, as explained in a science class.

Only a fraction of the potential causal understandings associated with categories and their members may be routinely invoked to constrain category-learning and use. Moreover, that fraction may show considerably more commonality across most people. Thus, there may be a skeletal set of causal patterns that most people use in similar ways, but that look little like fully detailed theories. However, we often mistake these sketchy relationships for richer intuitive theories because of certain illusions about what we really know.

BLESSED IGNORANCE

One way to better understand the importance of intuitive theories is to examine more closely the level of causal information that people normally track in the world around them. Surprisingly, the grain at which causal information is encoded has rarely been examined directly, with the result that the literature contains a large range of views ranging from those who argue that we have rich and powerful naïve theories (e.g., Vosniadu & Brewer, 1992) to those who argue that we detect only the weakest fragments of real world patterns (e.g., di Sessa, 1993). Thus, it seems clear that some data reduction must occur in our tracking of causal relations in the world around us; the question is just how much reduction and of what type.

One cannot exhaustively assess all causal patterns that people notice both because there are too many domains to examine and because we have no easy way to quantify the full range of what might be known and compare it to what is known. We can, however, ask how well people’s first impressions of what information they know corresponds to what they really know in local domains. We have conducted a series of studies examining people’s initial self-assessments of their knowledge and their re-assessments after a series of experimental manipulations. Those studies focus on judgements about knowledge of explanatory relations and contrast them to judgements about knowledge in several other domains, most notably knowledge about procedures, narratives, and facts (Rozenblit & Keil, 2002).

In tasks assessing explanatory understanding, people are presented with a large list of phenomena and devices and are asked to judge how well they

think they understand each of them. For example, they might be asked to judge how well they think they know how a helicopter works. Before they start rating their understanding, they are trained extensively on the use of a seven-point rating scale showing them that “level 1” would barely go beyond knowing the phenomenal properties of the device or phenomena (e.g., helicopters are things that fly up and down as well as sideways and use big blades on top and do not have wings), while “level 7” would be a fully detailed mechanistic understanding (e.g., a full description of the workings of a helicopter that captures the workings of all its parts).

After people have been trained on the scale, they rate a large set of initial items. Next, they give responses for a small subset of the items from that initial list. For each item there are several stages of responses. First, the participants are asked to write out the fullest description they can of how the device works or why the phenomenon is the way it is. Then, based on their self-perceived success of writing that explanation, they are asked to re-rate their initial understanding. Next, they are asked a critical diagnostic question requiring deep understanding (e.g., tell me how a helicopter goes from hovering to flying forward). They are subsequently asked to again re-rate their initial understanding in light of their answer to the diagnostic question. Finally, they are presented with a concise expert explanation of how the device works or why the phenomenon exists and are asked once again to re-rate their initial understanding.

People in these tasks consistently show a large drop in the ratings of their own knowledge after seeing an expert description, often with strong emotional reactions. They are often astounded at how little they knew compared to what they thought they knew. They do not say that they had misunderstood the scale; they are genuinely taken aback at how badly they overestimated their knowledge. The drops in ratings are found in several different populations and in different task manipulations, including one in which participants know they will be asked the follow-up questions.

The illusion of knowing is not seen for many other kinds of knowledge. Ask people to estimate how well they know certain facts, such as the capitals of countries, or certain procedures, such as how to make an international phone call, or certain narratives, such as the plot of a popular movie, and they are usually well calibrated. We have conducted a series of studies in the domains of facts, procedures, and narratives, and we repeatedly find either a much smaller drop in ratings or no drop whatsoever in comparison to explanatory understanding. Because the illusion of knowing is particularly powerful for explanatory knowledge, as opposed to knowledge of other types, we have called it the “illusion of explanatory depth” or IOED, meaning that people think they understand how things work and why phenomena exist in far greater depth than they actually do.

There seem to be several reasons why the IOED is so strong for causal information relative to other kinds of knowledge. One factor is a confusion of information that is stored in the head with information that can be recovered from phenomena that are in front of an observer. I may think I fully understand how a bicycle derailleur works because, when one is in front of me, I can puzzle out in real-time what all its parts do and why. But I may have mentally represented only a small fraction of the working arrangements and will be completely unable to recreate them without the object in front of me. This failure is analogous to at least one sense of situated cognition, where the cognitive capacities of individuals are said to be heavily dependent on the contexts in which they are situated (Brown, Collins, & Duguid, 1989).

We have found evidence for this factor by engaging in analyses of where the miscalibrations are largest for explanations of devices. There is a large drop in self-ratings for devices relative to other domains, but within the set of devices, there is also considerable variation. To understand the causes of this variation, we asked judges to rate several properties relating to each device, including familiarity with the device, the total number of parts in the device, the ratio of visible to hidden parts, and the number of parts for which a judge knew specific names. The strongest predictors of the drops in self-ratings were the ratio of visible to hidden parts and the number of parts for which judges knew names. By contrast, neither judged-familiarity of the item nor the total number of judged-parts-predicted drops (Rozenblit & Keil, 2002). It seems that the more visible parts an object has, the more one is lulled into thinking one has remembered those parts and internalised their working relations. Thus, visual clues that might be indicative of better causal reasoning when an object is present are confused as being indicative of richer mental representations. Knowing names for parts also seems to create an impression that one knows how they work.

The confusion of internally represented information with environmentally available information is similar to an effect frequently noted in the change blindness literature, where an observer of a scene may recall strikingly few details of a scene just observed while being convinced that she has internally represented far more. Indeed, follow-up studies show that people have a powerful “change blindness blindness”, in which they are unaware of the limitations (Levin, Momen, Drivdahl, & Simons, 2000). The similarity may reflect a common error of underestimating the extent to which one revisits a scene or a device to extract further information as needed rather than storing it all initially. In practice, why try to store extensive details about scenes or devices if it is easy enough to examine them again for further information when needed?

Three other factors converge to create a strong IOED relative to other domains: the difficulty of self-testing the quality of one’s explanations, the

smaller likelihood of having given them in the past, and a tendency to confuse insights gained at one level of analysis with details at a lower level.

Self-testing can be quite difficult because the end-state of an adequate explanation is often unclear. In asking myself how a helicopter flies, I may evaluate some misconceptions as evidence for detailed knowledge because I do not realise they are misconceptions until I have to provide a full explanation or am asked to answer a critical question about the phenomenon.

People also rarely provide comprehensive explanations for most phenomena and devices around them and it is therefore difficult to examine one's past successes and failures. By contrast, one often has a larger knowledge base of past procedures one has successfully completed or of facts one has recalled, or even of stories one has told, all of which can aid self-ratings of the relevant knowledge.

The confusion of one level of analysis with a lower one is a third factor. Most complex systems consist of functional units that interact in lawful ways. These units, in turn, have subunits that interact in functional ways as well. People may confuse genuine insight at one level (e.g., a cylinder lock works because the key makes the bolt go in or out when you turn it) with insight at a lower level (e.g., the key pushes up a series of pins in such a way that their ends are aligned with the edge of a cylinder which is then free to turn, thus enabling rotation that can move a bolt in or out). Most natural and artificial systems have nested sets of such stable subassemblies (Simon, 1981), and the rush of insight that comes with the understanding of high-level functional units may be mistaken for having an understanding of lower level units. This structural property of systems, for which we make causal explanations, is not present for facts and appears to be considerably weaker for procedures and narratives. Many procedures, such as how to make an international phone call, are a chain of steps with few or no embedded sub-steps. While narratives can have hierarchical structures (Mandler & Johnson, 1977), one tends to tell them as a chain of events at the lowest level and may therefore inspect one's knowledge more reliably and consistently at that level. A related factor may be the confusion of high level functions with lower level mechanisms, suggesting that the illusion of knowing might not be as strong for explanatory knowledge in domains with no functional concepts, such as those of non-living natural phenomena like the tides or earthquakes. We have some suggestive evidence that the IOED is somewhat weaker in those cases in contrast to devices or living systems (Rozenblit & Keil, 2002).

The IOED is different from classic overconfidence effects. For example, in the judgement and decision-making tradition, the disparity between people's average confidence levels for their answers to almanac questions and the proportion of correct answers is used to calculate an over-

confidence measure (Fischhoff, 1982; Lichtenstein & Fischhoff, 1977; Yates, Lee, & Bush, 1997; Yates, Lee, & Shinotsuka, 1996). However, that method assesses people's estimates of their performance on a task, not their differences in self-ratings over time from first impressions to having already attempted to generate the knowledge, to after being given the correct knowledge. In addition, much of that overconfidence literature remains controversial because people are asked to make probability estimates about single events. When they are asked to make frequency judgements, in fact, overconfidence effects sometimes disappear (Gigerenzer, Hoffrage, & Kleinboelting, 1991). Asking someone how likely it is that they just made a correct judgement is very different from asking them about the quality and detail of their knowledge.

Overconfidence is also found in studies of text comprehension, in which people often do not detect their own failures to understand a passage of text (Glenberg & Epstein, 1985; Glenberg, Wilkinson, & Epstein, 1982; Lin & Zabrocky, 1998). These studies, however, have people assess knowledge that they have just learned. In contrast, our IOED studies examine long-standing knowledge that people bring with them into the laboratory. The IOED is also not a phenomenon confined to arrogant students in an elite university. Indeed, when broader populations are examined, if anything, the illusion seems stronger in less-educated participants (Rozenblit & Keil, 2002). (See also Krueger & Dunning, 1999 for a related finding.)

One ongoing study in our laboratory suggests another important contrast to overconfidence effects related to self-image. Several studies have documented a self-enhancement effect, in which most people think they are above average on most positive traits (e.g., Krueger, 1998; Paulhus, 1998). This effect is a logical impossibility that apparently arises from inflated estimates of one's own abilities relative to others. The ongoing study on the IOED is suggesting that no such self-other difference exists for judgements of the depth of explanatory understanding; that is, judges are just as miscalibrated in their estimates of the abilities of others to offer explanations as they are of themselves. Thus, the IOED patterns vary differently across the self-other divide than do most other self-ratings.

The selectivity of the IOED for explanatory forms of understanding also seems to be present throughout much of development. Using the same paradigm as with adults, but simplifying the language and the examples, it has been possible to show drops in self-ratings over time in young elementary school children with a comparable specificity to that seen in adults for explanations, as opposed to other kinds of knowledge, such as procedures (Mills, Skinner, Goldenberg, & Keil, 2001). Thus, the structural properties of explanatory understanding that create the strength and specificity of the IOED are already at work from an early age. Younger

children also show higher levels of confidence for all kinds of knowledge but the selective illusion for explanatory understanding remains.

Most laypeople, and presumably many cognitive scientists as well, assume that ordinary folks' intuitive theories are much richer and more detailed than they really are. Since most cognitive scientists have been vague about the details of intuitive theories it is harder to demonstrate their capture by the IOED. It seems likely, however, that cognitive scientists fall prey to the same biases as most other people. Presumably, no cognitive scientist thinks that the average person knows every one of the several hundred thousand components of a 747 jet and all their functional roles. Many researchers, however, by incorrectly assessing their own knowledge of jets, may assume significantly more detail than normally is present in laypeople.

ESSENCES—WHAT LIES BENEATH?

Perhaps all the work supposedly done by theories can really be done in shorthand by beliefs in underlying essences. Beliefs in essences are said to guide many cognitive activities of both adults and children (Medin & Ortony, 1989; Gelman, 1999; Keil, 1986; Braisby, Franks, & Hampton, 1996; Gelman, 2003). In this view, being an essentialist means positing unseen features and properties that are assumed to be more at the core or "essence" of what an entity is than what is available through direct inspection. These essentialist beliefs are said to have a placeholder function for essential entities that are not explicitly known but which are assumed at the core of a category. For some, this view raises a concern as to whether the essentialism bias captures anything different about category knowledge (e.g., Gärdenfors, in press; Malt, 1994). If it just encompasses more features, then those features, once known to a person, might work in exactly the same psychological manner as prior more apparent features, thus making any theory/feature frequency contrast irrelevant. The placeholder function of the essentialist bias could be relegated to a relatively minor role of saying that one's feature list is incomplete and that one should hedge one's bets in making category-based judgements.

If essences are simply understood as defining features, or what Gelman (2003) refers to as "sortal essences", then an essentialist bias might indeed be little more than a feature-weighting and hedging function. There is a difference sense, however, of "causal essentialism" (ibid.) that has much stronger implications for the nature of theory-like influences. Essences are not simply assumed to be defining features, but also the causal reason behind the manifestation of surface features. The essence of gold, whatever it may be, is assumed to be causally responsible for all the phenomenal properties of gold, and so also for tigers, roses, and all other

natural kinds. It is not necessary that beliefs in essences be correct (Medin, 1989). Indeed, essentialist biases about race (Hirschfeld, 1996), species (Wilson, 1999), and gender (Taylor, 1996) are almost certainly mistaken and detrimental (Gelman, 2003). The causal essentialist bias therefore attributes to children and adults alike not only the assumption that many categories have hidden essences but also the belief that those essences are the reason behind many of the features of a category. Interestingly, the causal essentialist bias does not usually include any sense of how it is that the essence is causally linked to the surface, just the notion that it is.

In many cases, however, a belief in essences may entail more than merely believing they are responsible for surface features. It may also include some sense of the kinds of unseen causal patterns that are responsible for surface properties. A belief in essences might also include cases in which there are no hidden features at all as part of the essentialist bias, but only hidden patterns. In such cases, the bias assumes that there is a pattern of causal relations between the features that is responsible for their presence and/or their stable occurrence together. One example might involve the principles that lead to the formation of a solar system. One might easily observe all the components of a solar system but assume a set of non-obvious causal relations that explain its stability and which represent its true essence. There is no inner “stuff” to point to as the essence, just particular patterns of causation.

A second example might occur in the sophisticated biologist’s concept of species. Understanding species as a fixed set of hidden properties, such as a specific DNA sequence, is not an option for biologists since many species are distributions of such sequences where quite possibly no two members of that species (if it is one that does not have monozygotic twinning) have the same DNA. The scientist could simply assume the essence is a family resemblance of DNA types or the scientist could also assume that a particular set of causal relations creates a complex of related DNA types that, while capable of drifting over time, has great stability relative to other DNA sequences that are not governed by those causal relations. The stability itself may be the essence of the species and indeed current biology contains fascinating debates about the relative roles of the causal patterns of evolution, development, and reproduction in weighting various sequences of DNA in decisions about species (Wilson, 1999).

With laypeople, for natural kinds at least, there may be more of a tendency to assume a set of fixed essential properties; but even, in those cases, it seems likely that those essential properties are understood in terms of their causal roles in creating and maintaining phenomenal properties. Laypeople may often think of hidden properties as the peak of an upside-down pyramid, where the base is the set of surface manifestations that arise from a complex matrix of causal forces that make up the

volume of the pyramid. It may be rare for an essence to be understood merely as invisible essential features without any concomitant idea not only of the essence's enormous causal influence but also of the ways it might have that influence. The implication here is that the essential feature must have a property that is plausibly causally connected to the surface features. If there is no reasonable causal pathway linking the essence to the surface, it will be ruled out, even if other data are equally supportive of it.

The nature of an essence varies considerably as a function of the kind of thing involved and people's invocations of the relevant causal patterns will be correspondingly different. For living kinds, it is relatively easy to envision a causal process internal to a species that is largely responsible for the creation of manifest surface properties. In contrast, some have argued that artifacts have no essence at all (Schwartz, 1977; Sloman & Malt, this issue); others maintain that their essence lies in the intentions of their creators (Bloom, 1996). Thus, if essences exist for artifacts, they are not a set of properties and causal relations inside the artifact; they are instead the external goals of intentional agents. But in many cases, a disembodied intention as essence may be inadequate and we may also invoke the set of causal forces that explain a consistent intention to create them. It may not be enough to have an intention to create *X* for something to be *X*. The intention may have to come about in a reasonable manner.

For example, imagine that Adam wanders into a surgical suite of his neighbourhood hospital and sees an array of surgical instruments lying on a table. He is particularly intrigued by one instrument, which has a label on it calling it a "re-seater" which he pockets and takes home. Adam is a skilled machinist and carefully duplicates the re-seater for sale on the black market. Adam's clear intention is to make a surgical tool, yet unfortunately for Adam, the object he copied happened to be a plumber's tool that was accidentally left on the table by a plumber who had just fixed a faucet in the surgical suite. We do not think the thing Adam created is a surgical tool despite his clear intention of doing so because the broader context that explains the intention suggests otherwise. In many cases, laypeople may embed notions of intention in such larger contexts when they make judgements about artifacts.

Similarly, in viewing a novel object that we assume is an artifact, we often attempt to divine the intention of the object's creator from inferences about its preferred function and then use that intention as the basis for categorisation. But that set of inferences may often contain assumptions about reasonable ways in which intentions give rise to artifacts and not be so compelled by implausible ways such as the unwitting copier of the plumbing tool. Much of this awaits empirical studies on people's intuitions of how intentions influence categorisation across different contexts that vary the causal roles the intentions play. Here I

want to raise the possibility that beliefs in essences usually involve some grasp of larger causal systems, even for artifacts.

Essences may therefore often presuppose much more causal relational structure than is obvious at first glance. To the extent that they do, they are not shorthand for intuitive causal theories but a reflection of them. They may rarely be pure placeholders. At the same time, this putative knowledge of causal relations is almost never that of detailed mechanism. What else might it be?

SHALLOWNESS AS A VIRTUE

The relevance of causal structure to essence only intensifies concerns about how the supposed influences of knowledge of causal structure can be reconciled with the IOED. The IOED suggests that our folk theories are much coarser than we think. Is there enough structure and substance left to those theories to enable them to have their supposed influences on concepts and categorisation?

There are many ways we can track causal patterns that occur far above the level of concrete mechanisms. In what follows I describe three such ways in which we do monitor causal relations, starting with the coarsest, causal relevance, followed by causal powers, followed by coding of high level interactions among stable subassemblies. I will then further argue that the information that we do successfully track in these cases has significant influences on categorisation.

Causal relevance

Coding of causal relevance does not encode specific patterns of causal interactions but rather a sense of what properties matter most in a particular domain. Consider, for example, encountering a novel artifact and being told it is a kind of hand tool. Despite its having a distinctive colour and pattern of surface markings, one is inclined to discount those properties in developing a concept of the category to which that tool belongs. In contrast, one is inclined to count quite heavily the shape of the tool and its size. With a plant, however, the colour and surface pattern might be seen as quite central to the category with size being somewhat less important. This kind of knowledge, called “causal relevance”, can be elicited in a variety of ways and yields distinctive profiles for high-level categories such as tools, furniture, animals, and plants (Keil, 1994; Keil et al., 1998). Causal relevance is information about what kinds of properties are likely to matter in a domain but does not specify precisely how those properties will matter or even which property in particular will matter. Thus, causal relevance may indicate that colour and surface texture are

important for understanding living kinds but may not specify which colour or texture in particular.

There are individual exceptions. For example, some tools, such as measuring tapes, have surface markings that are essential to their nature. But overall default expectations about causal relevance remain. We have demonstrated these expectations in a series of studies with adults and children. Thus, if one is learning about a novel tool, plant, or animal, one tends to weight different kinds of properties as more important in the learning process, even though one knows little about the details of how those properties work (*ibid.*). For example, if participants are told about a novel hand tool that has a certain colour, shape, size, and surface pattern but are provided no other details, they are inclined to categorise new tools that have the same shape and size but different colour and surface patterns as more likely members of that category than other new tools that have different shapes and sizes but similar colours and surface patterns. By contrast, for novel flowers, they are more likely to weight all dimensions about the same. Similarly, in an induction task, if taught that a particular novel tool has a certain property, such as a distinctive colour or shape, participants are more likely to induce that other members of the same category have the same shape than the same colour. The pattern of inductions results in a very different relevancy profile for novel biological kinds, where colour is projected much more strongly to other members of the same kind.

We have default expectations about what sorts of properties are likely to do important causal work in such broad domains as animals, plants, hand tools, and furniture. At lower levels, such as insects vs. mammals, an inventory of the most causally relevant properties reveals them to be essentially identical. Thus, to the extent that specific colour is considered causally important for most mammals, it is considered causally important for most insects. Similarly, colour is seen as the same in importance for most farm tools and most scientific instruments. In all cases, a person may think that a kind of property is important, such as colour for animals, but have no idea of its particular causal role.

Recently, we have explored an influence of causal relevancy intuitions on the judged quality of explanations. We have given several descriptions of pairs of people who both claim to be experts on a particular class of things. The descriptions contain vague statements of the kinds of properties the person thinks are critical to understanding the class of things. Thus, we might be told that there are two people who claim to know all about phlebots, which are a kind of surgical instrument. Person A says the most important things to know about phlebots are that they are mostly black, have diagonal stripes on much of their surface, and have 23 parts. Person B says that the most important things to know are that they

are typically about the size of a shoe, are crescent shaped, and are quite fragile. If asked who is more likely to be the real expert, adult judges strongly prefer person B for artifacts. When the phlebot is described as a kind of mammal participants either find both experts about equally compelling or prefer case A. Children as young as 5-years old show these same preferences. The judged quality of explanatory understanding therefore appears to be strongly influenced by one's general default assumptions about what properties are mostly likely to be causally central, with the expectation that good explanations will emphasise the more causally relevant features.

Causal relevancy normally has a different directionality for artifacts as opposed to natural kinds. The causally relevant properties for artifacts tend to be ones that have direct consequences for the use of the artifact. By contrast, with plants, for example, not all causally relevant properties are seen as having strong causal consequences; instead, properties are causally relevant because they are assumed to be indicative of core causal factors. For example, I may not attribute any causal role to the white colour of a certain mushroom or to its surface dot pattern, but I likely will assume that the colour and pattern are tightly causally linked to the chemical makeup of that mushroom and that a differently coloured and patterned mushroom is not likely to be of the same kind. Perhaps the most salient causal aspect of the mushroom is that it is highly poisonous. I may have no beliefs about its surface pattern causing it to be poisonous; but I may well believe that the genetic processes that give rise to its surface patterns are closely linked to those that give rise to its poisonous properties. For a group of naturally occurring mushrooms to systematically vary from the target in colour and surface pattern, I assume they probably also vary in deep ways that also cause variations in their poisons.

In short, sometimes a property like colour can play a direct causal role for a living kind, such as camouflage or mate attraction; but other times it can be seen as tightly causally linked to the entity's essence in such a way that it is implausible for a member of that kind to exist with a radically different colour and surface pattern while maintaining all other properties unchanged. It may well not be true for some living kinds, which can show tremendous variations of colour (such as some species of flowers); but there is a clear general bias to think that whether something is an animal, a plant, or an inorganic substance, its colour is more directly causally linked to its most important properties than it is for a hand tool, a piece of furniture, or a farm implement. The nature of its linkage, however, does not need to be specified by notions of causal relevancy. As seen earlier, a causally relevant property can either be one that has a causal impact of its own or is judged as tightly causally linked to other features that do have such impacts.

Tracking of such causal relevancy notions may be one of the most basic and primitive aspects of noticing causal patterns in the world. Not only are the same relevancy profiles found at a wide range of ages, recent work suggests something like such profiles in other primates, including cotton top tamarins, a group of New World monkeys with a very modest brain size that is only a small fraction of that found in humans. In those studies, the tamarins generalise tool concepts on the basis of shape relative to colour but do the opposite for classes of foodstuffs, putting more of an emphasis on colour (Santos, Hauser, & Spelke, 2001). Given such findings in other species, it is perhaps not surprising that this sort of knowledge is usually implicit. Thus, in our studies, after adults generate their causal relevancy profiles, in debriefing, they often note that they had never explicitly thought of such systematic relations between classes of properties and high-level classes of things.

Causal powers

A more detailed manner of tracking causal structure extends beyond noting that particular property types likely have important causal roles in a specific domain, to encompass notions of their particular roles. This is one sense of the notion of “causal powers” (Harré & Madden, 1975). Thus, I may know not only that shape is important to the class of artifacts known as boomerangs; I also may believe that their distinctive shape gives them the ability to return in flight. Similarly, I may know not only that colour is important for bears; I also believe that it helps conceal them as predators. This kind of knowledge may not contain any further explanations about causal roles. I know that magnets have the ability to exert an attractive force on various metals but may know little about magnetism and the reasons that some metals make good magnets while others do not. We can think of this level as the first level at which distinct causal roles are attributed to properties. There may be a relatively small set of causal relations such as: contain, prevent, support, and launch that are at this first level of coding. Above that level of understanding, we simply know that a property is causally relevant for a kind. Notions of causal powers do not require any interrelations among properties in a coherent system; they may simply be isolated causal attributions to kinds. Magnets have the power of attracting certain metals, chairs of supporting human agents, and knives of cutting. How they come to have these powers may remain unspecified. Beliefs in causal powers therefore need not include any sense of mechanism.

Causal relations

A final, somewhat richer, but still abstract, way of tracking causal patterns is in terms of functional relations among stable subassemblies (Simon,

1981). In terms of coarse encodings, the stable subassemblies with all their constituent causal processes are treated as single entities in causal pathways with no tracking of their internal processes. We do encounter new questions concerning the definition of stable subassemblies and whether certain kinds of stability are especially prone to being treated as non-decomposable units; but that set of issues seems tractable and an important way of understanding how we do handle complex causal information. Thus, if a set of elements forms a stable unit with a clear function, we are inclined to encode only the unit as a whole and its functions, even if its internal structure works at the same level of complexity. For example, in understanding a complex mechanical watch, all parts are explained in terms of simple mechanics, but in many cases it may be useful to focus on interactions among larger subassemblies, such as the oscillator, the mainspring, the escapement, and the display. Understanding complex systems only at the highest level of functional interactions can lead us into trouble in some difficult cases but works well enough much of the time.

Knowledge of the coarsest functional relations in a domain may amount to little more than knowing the function of an entity as a whole and just a few of its largest constituents. For many people, their mental representations of the causal relations for cars may largely be confined to knowing that they convey people from place to place on roads, that they are propelled by an engine whose output is increased by pressing on an accelerator, and that they are slowed down by brakes. For an unfamiliar vehicle, only the notion of transport and some means of controlling speed may be present.

Collectively, these coarse representations of the world's much richer causal structure play a major role in how we learn and use knowledge about categories. They guide our attention and weighting of features and consequently our identification of new members of categories (Ahn, Kim, Lassaline, & Dennis, 2000). This coarseness may be one reason, however, why there is so much controversy concerning the role of theories in guiding concept structure and use. If theories are to be thought of as detailed mechanistic models of the world around us, at most we know local fragments that may differ considerably from person to person. But those mechanistic fragments may not be the central factors that influence our concepts and categorisation. Instead, the more skeletal frameworks of understanding may be much more universal and invariant across people.

All three of the coarse interpretations of reality just described frequently seem to operate at an implicit level. In our laboratory studies, participants clearly track causal patterns but often are unaware of those patterns until their set of responses are shown to them. I have already noted that people have sharply contrasting causal relevancy profiles for animals, machines,

and non-living natural kinds; yet this understanding seems to be implicit unless it is explicitly pointed out. Similarly, causal powers are often sensed and not explicitly mentioned. Finally, the highest-level causal functional roles of objects are often grasped but not discussed. We often use this kind of knowledge as a lens to interpret reality and, in looking through this lens, are often unaware of the ways in which it guides us to track some sorts of causal relations more effectively than others. While detailed mechanistic understandings are usually explicit and verbalisable, much of the coarser ways of tracking the world seems to occur outside awareness. This implicit aspect of causal understanding may be a key reason why developmental patterns are often described in such different ways. If there is a focus on ability to provide explicit accounts of how things worked or why phenomena exist, younger children often seem to be incompetent. By contrast, if one considers their patterns of judgement and the information needed to drive those judgements, young children can be highly competent in being sensitive to high-level causal patterns associated with broad domains.

DEPENDENCY AND DEFERENCE

A problem with only having coarse encodings of causal structure is that, when pressed, one runs into huge gaps in knowledge. Most of the time these gaps do not bother us, for two reasons: either we do not notice them because of the IOED, or we do notice them but assume that someone to whom we can have access knows them. More than 25 years ago, this dependence on others was pointed out in Hilary Putnam's essay on the meaning of meaning, in which he invoked a "division of linguistic labour" to explain how we successfully use terms like *gold* without knowing anything at all about the atomic makeup of gold (Putnam, 1975). In short, we succeed because we believe there is a relevant group of experts to whom we can defer when we encounter gaps in our own knowledge. In the case of *gold*, Putnam argued that we believe experts have knowledge of the true essence of gold. Putnam's proposal, combined with that of others (e.g., Burge, 1979), has led to a vigorous debate in cognitive science over "narrow" versus "wide" content, that is, whether meanings can be individuated solely by referring to internal mental states or whether they are also bound to the external world (Segal, 2000; Fodor, 1998). Such a debate affects claims about the nature of concepts, but it is less relevant to questions about factors that influence categorisation. One might be unsure about whether concepts have theories as part of their structure, while still maintaining that certain kinds of causal explanatory relations influence categorisation. Similarly, one might maintain that the division of linguistic

labour influences judgement about word usage, while being unsure whether word meanings intrinsically depend on such a division.

The division of linguistic labour is a special case of a more general division of cognitive labour that occurs in all cultures. Just as we rely on specialisation of physical labour to provide us with resources that we cannot produce ourselves, we rely on a specialisation of intellectual labour to provide us with underpinnings to knowledge where we have none. Virtually every human group of any size develops different sub-communities that have different clusters of specialised knowledge and deeper explanatory understanding. But this reliance on other's knowledge is a far more subtle and complex ability than it appears at first and links in powerful ways to our tracking of causal patterns in the world.

Consider, for example, variations in knowledge about trees. I am profoundly ignorant about different types of trees and know little beyond the distinction between trees with needles and trees with leaves. Thus, when confronted with a particular tree and asked whether it is an elm, a beech, or a basswood tree, I would have no idea. Yet, I firmly believe that there are such categories and that there are people who could tell me why trees belong in each category and something about how they are related in a larger system of classification. Moreover, I do not necessarily believe that the best expert is simply someone who has looked at a lot of trees. There are hikers who may go through a wood almost every day for years and be highly experienced with various trees; but that experience may be with landmarks on a trail and not with trees as a species. The hikers may be able to recognise individual trees in those woods better than almost any one, but they may not have any sense of tree categories; or perhaps they have a sense of categories in terms of tree types that show tree blazes the most clearly. In deciding where to allocate my dependence on other's knowledge, I would put my trust in someone who I think could explain all the various surface properties of trees and their "behaviours" over the seasons. I would assume that a person could do so because she knew something about the deep relations between certain classes of trees.

To pick the right experts, I need to have some sense of biology and perhaps of plants as well. I need a skeletal sense of the kinds of relations that are central to a domain such that a person who grasped those relations would be much more likely to explain surface phenomena in the domain as well. To benefit from the division of cognitive labour, I need to have a sense of what the key principles are in broad domains, such as animals, plants, and tools, such that a person who grasped those principles would be able to guide me to proper judgements about categorisation.

Trees are an especially interesting case because they illustrate how one's assumption about the division of cognitive labour may be wrong. Thus, the global category of "trees" does not agree well with western biological

sciences, which would instead classify an apple tree as more similar to a daisy than to a pine tree. In turn, a pine tree would be seen as more similar to a fern than to an oak (Dupre, 1981). I may be correct in assuming there are molecular and evolutionary reasons that shed insight into the differences between elms and beeches but wrong in assuming that there are any molecular reasons for trees being a separate category from other plants. There are, however, evolutionary arguments that the appearance of tree-like structures is due to the biomechanical constraints of obtaining adequate light for large free-standing plants to survive (e.g., Niklas, 1996), though these accounts reveal no common molecular relations. There are yet other causal systems in which trees may be embedded that are quite different from those of western science but which may well have their own groups of experts who are especially tuned to those kinds of causal structures, e.g., trees that are particularly good hosts for certain kinds of fauna versus those that create environments that are especially hospitable to certain kinds of plants (López, Atran, Coley, Medin, & Smith, 1997). These alternatives could suggest a “promiscuous realism” in which there are an indefinitely large number of such categories, since they reflect the boundless nature of human creativity (Dupre, 1981). The alternative view favoured here allows many natural and artificial objects to be parts of several different stable causal systems but not an indefinitely large set. These systems could each have their own experts and ways of construing categories but would be limited to a relatively small number of real world stable causal systems that embed those kinds (Keil, 1989).

By this account, categorisation may be heavily influenced by causal interpretations that not only tell us what properties and kinds of relations are likely to be relevant, but also what kinds of experts could provide us with further details. For such an account to be plausible, however, it is important to show that people have reliable intuitions about the division of cognitive labour that at least show some consistency within cultural groups. In a recent series of studies, we have explored such intuitions by asking both adults and children to tell us if a person who understood phenomenon A was more likely to understand phenomenon B or C. This is a very natural task even for preschoolers (Lutz & Keil, 2002). In adults, the task is most sensitive when set up as a triad of the following sort:

John knows all about why gasoline is poisonous to people. Because of this, what else is he likely to know a lot about?

Why horses perspire when they get hot, or

Why a heavy person must sit closer to the middle on a seesaw.

Both adults and children will pick the alternative about perspiration most often, even though the first and third sentences refer to people whereas the second refers to horses. They often do not know the details of

the answers at all, but can nonetheless be quite confident in their judgement. If adults are asked to provide a rationale, many say that the person is a biology expert. Others who have more difficulty voicing a reason are still confident in their choice. Children, however, often cannot give a reason, even though they show the same clusterings of biology with biology and physical mechanics with physical mechanics. When children do offer justifications, they usually refer to different relational patterns. For example, one child said, "John knows about how people and animals work, what their insides do" "... the other thing is about how things move".

The knowledge that drives these judgements creates distinctions like many of the major natural and social science departments in universities: physics, chemistry, biology, psychology, political science, and economics. These judgements happen in younger children who have no awareness of these labels or the departments. With adults and older children, further subdivisions such as molecular biology vs. ecology are also understood without explicit labels for those disciplines. Thus, people have a sense of what sorts of phenomena are likely to be clustered together because they can be explained by some common set of principles. A person who really understood one phenomenon in that cluster would also tend to understand others that arose from the same principles, even if they were radically different on the surface

Laypeople in fact do not intuitively know the principles of modern physics, chemistry, or biology but instead have more schematic notions that approximate the domains of a science. Physics is understood as being about moving solid objects and their interactions with other solids, chemistry as the ways objects change state or mix with others, and biology as the basic functions of living kinds. It is possible to reveal these approximations by presenting phenomena that are technically problems in physics, chemistry, or biology but which may elude children and many adults because they do not fit these simple schemata. For example, if a child's schema dictates that physics involves bounded objects in motion, a phenomenon involving static forces, such as those holding a suspension bridge, may not be clustered with other physics problem.

These studies on the division of cognitive labour suggest that young children link together phenomena that they think are governed by the same causal patterns with particular groups of experts. Even for preschoolers, it is very natural to see such groups of experts as mapping onto causal patterns in the world. Put differently, to solve the above problems, children and adults alike need to have some sense of how the world is causally structured into different domains. They do not need to know the details of how things work in each domain, such as the principles of respiration or reproduction for biology; they merely need to have enough of a sense of the causal patterns distinctive to that domain. That

skeletal sense of causal patterns may well be the same as the coarse level of representing causal relations that was discussed earlier. In short, people do track large-scale causal patterns in coarse terms that may explain both theory-like influences on categorisation and intuitions about how causal understandings are clustered in other minds.

THE COGNITIVE CONSEQUENCES OF KNOWING WHO KNOWS WHAT

If children as young as 4 years have reliable notions of how knowledge is distributed in other minds, and if several of those notions are based on their tracking of high-level causal patterns, what do people do with this knowledge? In particular, to what extent do implicit models of the division of cognitive labour influence categorisation? One influence may involve arbitration of marginal cases or the enabling of conceptual change when an initial category structure is missing key relations. In such cases, people often show deference to experts by adjusting their categories when told that an expert has a particular view of category membership. The extent of this deference, however, may vary across contexts and may not always include the most appropriate uses of experts. Thus, in some cases, people might defer on a natural kind categorisation decision to a group described as shoppers as much as they do to a group described as scientists (Braisby, 2001). The opinions of both experts and non-experts do matter to our decisions about category membership but perhaps not in a simple manner that has the opinions of the most relevant scientists always being the most influential. In addition, if the views of the experts are discordant enough with all other known cases of expert influence (e.g., a group of experts who claim that cats are really remotely controlled robots) deference will not be nearly as strong as with a more mundane but more ordinary case (e.g., an iris is really a kind of orchid).

Adults are not usually confronted with an expert opinion that causes them to radically change category assignments. When we do make revisions they are most often to quite nearby categories. Even with nearby categories, wholesale reassignments are rare, e.g., the discovery that panda bears were not really bears. Larger reassignments may be more common in children, e.g., learning that whales are not fish, but such dramatic revisions may not be the norm even in childhood.

Notions of the division of cognitive labour normally work in a more subtle and incremental manner during category learning. Consider how an adult might learn about a new category, such as a new disease agent. He might have heard the term “prion” mentioned a few times in relation to mad cow disease. As he learns more about prions and the disease, he needs to weigh different bits of information that he encounters, ranging from the

panicked remarks of a caller to a radio talk show to the remarks of a molecular biologist. If the caller and the biologist both ascribe a property to prions, the caller likely will weigh the biologist's ascription more heavily. New information about a category that comes from more qualified sources is more likely to be given more weight. The process is, of course, fallible, and urban myths about any number of categories arise from such fallibilities; but there is a general effect of favouring knowledge from appropriate experts. The relevant expert also clearly varies as a function of the kinds of categories and relations involved. I weigh the molecular biologist's statements heavily in gathering information about prions, but when the biologist starts talking about economic recessions brought about by diseases, I weigh that information less heavily in developing knowledge of a category of recessions.

One can therefore think of the division of cognitive labour as providing a spotlight on the most relevant features when thinking about a category. It can serve both to make features salient that might not have been otherwise noticed and to weight salient ones more heavily as causally central. These influences are not confined to the occasional novel category and, in fact, are at work for some of our most mundane and familiar categories. Consider, for example, my understanding of the category of dogs. For years I have noticed statements about the similarities and differences between dogs and wolves, but until recently these statements had been made mostly by dog owners, people with dog phobias, and authors of various novels involving ferocious dogs. Because I regarded all those sources as non-experts, I had not used them much to adjust my understanding of dogs. A few months ago, however, I came across an article on the evolution of dogs, in which biologists discussed the relatively short time frame in which dogs have emerged from the wolf category and the extraordinary overlap in their genetic material. That article caused me to weigh somewhat differently many of the features of dogs and their causal roles. I may have been mistaken in making such adjustments, but they occurred because I believed the information came from credible experts. Such a re-weighting does not cause me to label major groups of dogs, such as Labradors and poodles, differently, but they may influence cases at the margin. Thus, if I see a creature with wolf-like and dog-like features, I would be more likely to accept a more wolf-like creature as a dog.

Children's categorisation may be just as heavily influenced by their own quite well developed senses of the division of cognitive labour. As they learn about new categories or elaborate on recently acquired ones, they might well weight information differently based on its sources. In an ongoing line of research in our lab, we are finding that elementary school children discount or favour the same information as a function of who

provided the information. Thus, to the extent that information about a new category is learned through social transmission, the ways that information affects categorisation will be influenced by our understandings of the division of cognitive labour.

The impact of a sense of the division of cognitive labour is not just on information acquired through social transmission; it can work directly on information acquired from real time, direct experience. If some prior encounter with a body of expertise causes one to weight certain kinds of properties or causal relations as more central, that effect will carry forth to new encounters with potential instances of a category. In my own case, the features and causal relations that I encounter in novel canines will be encoded differently because of my beliefs about prior information that came from experts vs. novices.

In many cases, there is a cycle of interaction between notions of the division of cognitive labour, our tracking of causal structures, and the impact on categorisation. When encountering a novel phenomenon, I will notice certain high level causal patterns, such as those of relevancy, causal power, and schematic patterns, and will use them to pick out a relevant domain of expertise. Identifying the phenomenon as in the domain of biology will lead me to further consider what I have heard from experts in that domain in terms of key relations and properties. That information, in turn, will guide my more detailed analysis of the phenomenon and my attempts to form relevant categories.

A division of labour framing of information seems to promote searches for deeper causal relations. Thus, in a series of studies (Keil & Rozenblit, 1997), we compared adult ratings of the similarities of various phenomena, such as “water is transparent to light”, “water is a frequent source conflict between nations”, and “televisions get static buildup on their screen” in cases where they were presented as “bare facts” with cases where they were embedded in a division of cognitive labour frame (e.g., “This expert knows all about why water is transparent to light”). Adults categorised the phenomena quite differently when presented in their raw form versus when embedded in a division of cognitive labour frame in which clustering by similar experts is requested. In the explanation frame case adults see a much stronger similarity between the two cases that share more similar underlying causal patterns (in this case those of physics). Embedding phenomena in a frame that invokes the division of cognitive labour increases sensitivity to their underlying causal patterns and principles. Moreover, preliminary findings from an ongoing study with children suggest that this division of cognitive labour framing causes a corresponding shift in their similarity judgements.

Our sense of the division of cognitive labour also allows us to be more confident in our understandings when there are large gaps in our

knowledge. Put differently, that sense tells us what sorts of causal patterns and properties are likely to be explanatorily relevant in a domain and, at the same time, likely to be known in much more detail by experts. We can therefore be more confident if we believe that experts would be as well. Thus, I weight some properties more heavily because I both sense their important roles in causally central relations in that domain and because I believe that relations of that sort drive successful expertise. The division of cognitive labour helps give us a sense of relevant properties and relations within a domain. It may be analogous to the role of a guardrail on a narrow, curving mountain road. Drivers on such roads very rarely touch the guardrails, but most feel vastly more confident and willing to drive on roads with guardrails as opposed to those without and may use the guard rail to guide their driving speed and vigilance. The division of cognitive labour is a comparable guiding and supporting backdrop for categorisation.

CONCLUSIONS

When we leave laboratory categorisation tasks, which have a few neatly defined features, we give up the elegance of clear control over our stimuli. However, we then start to confront one of the most basic problems of real world categorisation: the immensity of information that is associated with most categories in our daily lives. Many features and feature correlations help create this immensity; however, another major contributor is the massive set of causal patterns, which are responsible for the creation and continued existence of members within a category and are critical to understanding which features are likely to be most central to that category. Fortunately, much of the time, typicality, correlation, and causation converge. The most frequent two features are usually highly correlated and usually play important causal roles for members of a category. Furthermore, patterns of variation of property types and values and property kinds within and across categories can be powerful clues to causal relations.

There are also cases of mismatch, in which causality and typicality are not correlated. For many artifacts, highly typical colours can be such properties, such as washing machines being white. For natural kinds, highly typical but low causally central features seem less common because one assumes an efficiency in which functionally irrelevant features are discarded. Indeed, for biological kinds, there may even be a bias to attribute important causal roles to structures when in fact none exist and they are mere byproducts of other structures (Gould & Lewontin, 1979). Still, even in biology, some features may be clearly irrelevant even if they are consistent with all known instances. For instance, I recently moved to a

neighbourhood where a parasite had attacked a special kind of evergreen I had not seen before. Virtually every tree of this kind has huge blotches of brown needles instead of a uniform green. I did not, however, attribute those properties as important to that class of trees, presumably because of beliefs about normal structure/function relations in plants.

An appreciation of causal relations does appear to be essential to understand categorisation in the real world. That appreciation, however, is not in the form of well-developed theories that provide blueprints of various devices and phenomena. One's knowledge does not come close to allowing one to recreate a working system, only to appreciating some central causal patterns that are collectively unique to a broad domain in which that object is a member. This coarse level of encoding is powerful enough to narrow down the complexity of what one must track, but also shallow enough to allow quick and efficient processing. Shallowness is a real virtue in this sense of navigating the causal complexities of the world around us. Thus, we extract the causal "gist" to ascertain enough detail within a particular domain so that we can detect the most salient features without being overwhelmed. In this view, the relevance of causal understandings to concepts and categorisation is a basic aspect of our cognition seen as early as humans can notice causal patterns for large-scale domains, something that quite young infants notice for such domains as intentional versus inanimate agents. What develops is an appreciation of ever-finer patterns and better ways of linking cause to typicality and correlation.

Causal information is valuable to understanding classes of objects around us, both in terms of predicting patterns and in terms of encoding relations and helping guide exploration. But, causal information threatens to swamp us with its complexity. It has been popular to invoke intuitive theories as a mechanism for significantly reducing the number of features and correlations we have to examine. However, this claim does not help us understand a two-faceted problem for theories: Which of the countless theories for a set of relations do we pick, and at what level of detail?

There has been little attention given to the problem of finding the minimal amount of information about a theory required to "get by" in tasks such as categorisation and induction. Our studies on the Illusion of Explanatory Depth (IOED) show that adults and children alike have a particularly strong illusion of knowing explanatory relations in far more detail than they really do. Moreover, this illusion is a distinct phenomenon from other overconfidence effects and is much stronger for explanatory knowledge as opposed to several other types.

One should not interpret the IOED studies as supporting the idea that intuitive theories of concepts and categorisation are too ephemeral and

sketchy to be of any use. Rather, we effectively track high-level coarse causal patterns, which tell us what sorts of causal relations are central to a given domain. That level of causal interpretation, however, is very different from concrete mental models of how things work and rarely includes notion of specific mechanisms. Indeed, it is often implicit and works outside of normal awareness and discourse about phenomena and devices. Such coarse representations become evident when one specifically looks for people's abilities to track causal patterns and explores how that information can guide explanation preferences and notions about the division of cognitive labour. I have also suggested that coarse understandings guide category-learning and use and may be the real basis for many theory-like effects on categorisation. I have further argued that our sensitivity to causal patterns is often heightened when we consider phenomena from the viewpoint of how such patterns might map onto domains of expertise.

A final note concerns the generative nature of explanatory understanding. One major factor creating the IOED is confusion between what one mentally represents and what is decipherable from a pattern that is present for inspection. People are often quite adept at figuring out causal relations and patterns on the fly when in information-rich environments. Thus, I may not store in my head detailed theories of desert, arctic, and jungle vehicles but, when confronted a series of vehicles and environments, I may quickly sort them into categories in which their feature clusters mesh nicely with these three environments. I come to the situation with schematic expectations about causal patterns in artifact domains and perhaps for vehicles as well, and I use those to help create a much more transient, detailed theory on the fly. Like "ad hoc categories" (Barsalou, 1983), much of the detail of our everyday theories may be fleeting and controlled primarily by local and, immediate contexts.

It may be more appropriate to think of what a person brings to a situation, not as involving just broad causal gists, but also involving a set of specialised toolboxes for constructing theories. That is, the plumber, the electrician, and the carpenter bring quite different sets of tools when making house calls, tools that embody expectations about the kinds of problems that will be encountered and which are designed to be most causally effective for those situations. We also tend to bring a conceptual toolbox that best embodies expectations about the most relevant kinds of causal patterns for each particular domain of phenomena. Such a toolbox would contain schemata that are most likely to be central to that domain and perhaps some information on how they work together in a larger system. In this manner, rapidly constructed ad hoc causal explanations may also show theory-like effects on categorisation that supersede broader, more abstract expectations.

REFERENCES

- Ahn, W., & Kalish, C. (2000). The role of mechanism beliefs in causal reasoning. In F.C. Keil & R.A. Wilson (Eds.), *Cognition and explanation* (pp. 199–225). Boston: MIT Press.
- Ahn, W., Kim, N.S., Lassaline, M.E., & Dennis, M.J. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology*, *41*, 361–416.
- Barrett, S.E., Abdi, H., Murphy, G.L., & Gallagher, J.M. (1993). Theory-based correlations and their role in children's concepts. *Child Development*, *64*, 1595–1616.
- Barsalou, L.W. (1983). Ad hoc categories. *Memory and Cognition*, *11*, 211–227.
- Bloom, P. (1996). Intention, history, and artifact concepts. *Cognition*, *60*, 1–29.
- Braisby, N.B. (2001). *Deference in categorization: Evidence for a division of linguistic labor?* Paper presented at the 2001 Congress of the European Society for Philosophy and Psychology, August 8–11, 2001. Fribourg, Switzerland.
- Braisby, N., Franks, B., & Hampton, J. (1996). Essentialism, word use, and concepts. *Cognition*, *59*, 247–274.
- Brown, J.S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, *18*, 32–42.
- Burge, T. (1979). Individualism and the mental. In P. French (Ed.), *Midwest studies in philosophy IV: Studies in metaphysics* (Vol. 4, pp. 73–122). Minneapolis: University of Minnesota Press.
- Carey, S. (1991). Knowledge acquisition: Enrichment or conceptual change? In S. Carey & R. Gelman (Eds.), *The epigenesis of mind: Essays on biology and cognition* (pp. 257–291). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- di Sessa, A. (1993). Towards an epistemology of physics. *Cognition and Instruction*, *10*, 105–225.
- Dupre, J. (1981). Natural kinds and biological taxa. *Philosophical Review*, *90*, 66–90.
- Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.) *Judgement under uncertainty: Heuristics and biases* (pp. 422–444). Cambridge: Cambridge University Press.
- Fodor, J.A. (1998). *Concepts: Where cognitive science went wrong*. New York: Oxford University Press.
- Gärdenfors, P. (in press). Concept modeling, essential properties, and similarity spaces. *Behavioral and Brain Sciences*.
- Gelman, S.A. (1999). *Essentialism*. Retrieved August 1, 2002 from MIT Encyclopedia of the Cognitive Sciences site <http://cognet.mit.edu/MITECS/Entry/kornblith>.
- Gelman, S.A. (2003). *The essential child*. London: Oxford University Press.
- Gigerenzer, G., Hoffrage, U., & Kleinboelting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*, 506–528.
- Glenberg, A.M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 702–718.
- Glenberg, A.M., Wilkinson, A.C., & Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. *Memory and Cognition*, *10*, 597–602.
- Gopnik, A.A., & Wellman, H.M. (1994). The theory theory. In L.A. Hirschfeld & S.A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 257–293). Cambridge: Cambridge University Press.
- Gould, S.J., & Lewontin, R.C. (1979). The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London B*, *205*, 581–598.
- Hampton, J.A. (2001). The role of similarity in natural categorization. In U. Hahn & M. Ramscar (Eds.), *Similarity and categorization*. Oxford: Oxford University Press.
- Harré, R., & Madden, E.H. (1975). *Causal powers*. Oxford: Blackwell.

- Hirschfeld, L.A. (1996). *Race in the making: Cognition, culture, and the child's construction of human kinds*. Boston: MIT Press.
- Keil, F.C. (1986). The acquisition of natural kind and artifact terms. In W. Demopoulos & A. Marras (Eds.), *Language learning and concept acquisition: Foundational issues* (pp. 133–153). Norwood, NJ: Ablex.
- Keil, F.C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: Bradford Books/MIT Press.
- Keil, F.C. (1994). Explanation based constraints on the acquisition of word meaning. *Lingua*, 92, 169–196.
- Keil, F.C. & Rozenblit, L. (1997). *Knowing who knows what*. Paper presented at the 1997 meeting of the Psychonomics Society, Philadelphia.
- Keil, F.C., Smith, C.S., Simons, D.J., & Levin, D.T. (1998). Two dogmas of conceptual empiricism. *Cognition*, 65, 103–135.
- Kuhn, T.S. (1977). *The essential tension: Selected studies in scientific tradition and change*. Chicago, IL: University of Chicago Press.
- Krueger, J. (1998). Enhancement bias in the description of self and others. *Personality and Social Psychology Bulletin*, 24, 505–516.
- Krueger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77, 1121–1134.
- Levin, D.T., Momen, N., Drivdahl, S.B., & Simons, D.J. (2000). Change blindness blindness: The metacognitive error of overestimating change-detection ability. *Visual Cognition*, 7, 397–412.
- Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Decision Processes*, 20, 159–183.
- Lin, E.L., & Murphy, G.L. (1997). The effects of background knowledge on object categorization and part detection. *Journal of Experimental Psychology: Human Perception and Performance*, 23, 1153–1169.
- Lin, L., & Zabrocky, K.M. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology*, 23, 345–391.
- López, A., Atran, S., Coley, J.D., Medin, D.L., & Smith, E.E. (1997). The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. *Cognitive Psychology*, 32, 251–295.
- Lutz, D.J., & Keil, F.C. (2002). Early understanding of the division of cognitive labor. *Child Development*, 73, 1073–1084.
- Lycan, W.G. (2002). Explanation and epistemology. In Paul Moser (Ed.) *The Oxford handbook of epistemology*. Oxford: Oxford University Press.
- Malt, B.C. (1994). Water is not H₂O. *Cognitive Psychology*, 27, 41–70.
- Mandler, J.M., & Johnson, N. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology*, 9, 111–151.
- Medin, D.L. (1989). Concepts and conceptual structure. *American Psychologist*, 44, 1469–1481.
- Medin D.L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning*. Cambridge: Cambridge University Press.
- Medin, D.L., & Shoben, E.J. (1988). Context and structure in conceptual combination. *Cognitive Psychology*, 20, 158–190.
- Mills, C., Skinner, H., Goldenberg, D., & Keil, F. (2001). *Thinking you know more than you do: Children's assessment of their own knowledge and explanations*. Poster presented at 2001 Cognitive Development Society Conference, October 26–27. Virginia Beach, VA.
- Murphy, G.L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

- Murphy, G.L., & Allopenna, P.D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 904–919.
- Murphy, G.L. & Medin, D.L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*, 289–316.
- Niklas, K.J. (1996). How to build a tree. *Natural History*, *105*, 48–52.
- Paulhus, D.L. (1998). Interpersonal adaptiveness of trait self-enhancement: A mixed blessing? *Journal of Personality and Social Psychology*, *74*, 1197–1208.
- Prinz, J.J. (2002). *Furnishing the mind: Concepts and their perceptual basis*. Cambridge, MA: MIT Press.
- Putnam, H. (1975). The meaning of “meaning”. In K. Gunderson (Ed.), *Language, mind, and knowledge* (Vol. 2, pp. 131–193). Minneapolis: University of Minnesota Press.
- Rehder, B., & Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology: General*, *130*, 323–360.
- Rozenblit, L.R. & Keil, F.C. (2002) The misunderstood limits of folk science: An illusion of explanatory depth, *Cognitive Science*, *26*, 521–562.
- Santos, L.R., Hauser, M.D., & Spelke, E.S. (2001). The representation of different domains of knowledge in human and non-human primates: Artifactual and food kinds. In M. Beckoff, C. Allen, & G. Burghardt (Eds.) *The cognitive animal*. Cambridge, MA: MIT Press.
- Schwartz, S. (1977). *Introduction to naming, necessity, and natural kinds*. Ithaca, NY: Cornell University Press.
- Segal, G.M.A. (2000). *A slim book about narrow content*. Cambridge, MA: Bradford Books/MIT Press.
- Simon, H.A. (1981). *The sciences of the artificial*, 2nd edn. Cambridge, MA: MIT Press.
- Sloman, S.A., Love, B.C., & Ahn, W. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, *22*, 189–228.
- Sloman, S.A., & Malt, B.C. (this issue). Artifacts are not ascribed essences, nor are they treated as belonging to kinds. *Language and Cognitive Processes*, *18*, 563–582
- Smith, E.E., & Medin, D.L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.
- Taylor, M. (1996). The development of children’s beliefs about social and biological aspects of gender differences. *Child Development*, *67*, 1555–1571.
- Vosniadou, S., & Brewer, W.F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology*, *24*, 535–585.
- Walker, J. (1979). The amateur scientist. *Scientific American*, *240*, 162–172.
- Wilson, R.A. (1999). *Species: New interdisciplinary essays*. Cambridge, MA: MIT Press.
- Wisniewski, E.J., & Medin, D.L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science*, *18*, 221–281.
- Yates, J.F., Lee, J.W., & Bush, J.G. (1997). General knowledge overconfidence: Cross-national variations, response style, and “reality”. *Organizational Behavior and Human Decision Processes*, *70*, 87–94.
- Yates, J.F., Lee, J.W., & Shinotsuka, H. (1996). Beliefs about overconfidence, including its cross-national variation. *Organizational Behavior and Human Decision Processes*, *65*, 138–147.