

# Cognition and Explanation

## Foreword

Much of human discourse, whether it be in a scientific meeting or in gossip, contains explanations. We all seem to need explanations and want to share them with others. Conversations without any explanatory components would strike us as the most banal sorts of cocktail party banter. Indeed, it has been claimed that those afflicted with Williams Syndrome may have just a deficit, being able to engage in conversations that report facts or follow practiced scripts well, but having much more difficulty creating or following explanations (Carey et al., 1993; Udwin and Yule 1991). It seems that explanations are such a fundamental part of our everyday cognition that their absence in otherwise normal discourse is striking and even suggestive of a neuropsychological deficit. What makes explanation so predominant in human life? What roles do explanations play? How do explanations function in the social world and in the lives of individuals? What is the nature of a good explanation? What is the relationship between explanation and mechanism?

Providing explanations seems to be a peculiarly human activity, and understanding them a specifically human ability. But, like other peculiarly human abilities, the most obvious being that of understanding natural language, this ability is not one we are able to deploy at birth, and it is unlikely to emerge wholly anew and be distinct from capacities that other species have. We develop our ability to provide and understand explanations through childhood, but how? What is crucial to this development? Is there always a need for an explanatory component to help organize knowledge even in the youngest child, or is it a later emerging part of knowledge that requires a great deal of prior non-explanatory groundwork? To what extent is the increasing ability to provide and understand explanations a natural part of growth of expertise in a specific domain, as opposed to the development of a domain general capacity?

In evolutionary terms, how unique is our ability to provide and understand explanations? If we think that explanations are essentially linguistic, and consider ourselves as the sole linguistic species, then we are likely to think of our abilities here as unique among species, and perhaps inaccessible even to preverbal human infants. But should we think of the tie between explanation and language in this way? Is one a precursor to the other, and, if so, which one and why? If we do think of explanation and language as intimately connected, then what precursors to the emergence of explanatory abilities exist in the rest of the animal kingdom? In what ways are *social* aspects important to the evolution and development of these

abilities? We might think that cognitive ethology provides a significant window on OUR explanatory abilities, even supposing our uniqueness here in the world of living things. And if we can shed light on that uniqueness by examining natural creatures that don't explain, how about artificial creatures that might explain?

As we raise these broad but still largely empirical questions about explanatory abilities, we see more philosophical issues and questions come to the fore. To move immediately to the most basic of these, what is an explanation? Should we think that there is any *one* thing here, explanation, that we can provide a unified account of? Or is "explanation" just an umbrella term whose particular instances bear nothing more than a family resemblance to one another? Philosophers of science have addressed this issue, in part, through exploring the question of how the various sciences are *unified*, for it is plausible to think that one goal shared by all sciences is that of providing explanations for phenomena in their domain. Do scientific explanations take a standard form, or can they be rationally reconstructed to do so, as the standard deductive-nomological model of scientific explanation supposes? How does one's view of the unity of scientific explanation interact with what we might think of as more globalist views of the unity of explanation? Does accepting or rejecting a unificationist view of scientific explanation make more or less plausible the idea that explanations constitute some type of kind?

To move back towards the type of psychological questions that are at the core of this issue, we can consider the question of whether there are *domains* of explanation. At one extreme, we might think that there are many diverse and distinct domains in which explanations operate. There is a *social* domain, where our "folk psychological" explanations are at home; there is a *physical* domain, about which we might have both naive and sophisticated theories; there is a *religious* domain with its own types of explanatory goals and standards, and so on. At the other extreme, we might think that these domains are not all that diverse, and that there are certain relations of dependence between them. For example, some have proposed that children are endowed with two distinct modes of explanation that come to shape all other types of explanation that children come to accept: an intuitive psychology, and an intuitive physical mechanics (Carey 1985). Clearly, we can move between more abstract, philosophical considerations of this issue to those that are to be answered through the empirical evidence provided by cognitive developmentalists – as we have just done in this paragraph.

One final issue concerns the role of the world in general and *causation* in particular in explanation and our attempts to make sense of our explanatory abilities. At the turn of the century, Charles Sanders Peirce argued that induction about the natural world could not succeed without "animal instincts for guessing right". It seems that somehow the human mind can grasp enough about the causal structure of the world to allow us to guess well. We know from the problem of induction, particularly in the form of the so-called new riddle of induction made famous by Nelson Goodman, that the power of brute, enumerative induction is limited. The idea that we and other species have evolved biases that enable us to grasp aspects of

the causal structure of the world seems irresistible. But which of these biases make for explanatory abilities that work? For explanatory abilities that get at the truth about the world? What roles, if any, do causal powers play in the explanations we develop? Do explanatory devices, of which we are a paradigm, require a sensitivity to the causal patterns that really exist in the world in order to be successful?

Many of the questions we have just raised are some of the most difficult in all of cognitive science, and we surely do not presume that they will be answered or even addressed in the articles that follow. We raise them here, however, to make clear just how central explanation is to cognitive science and all its constituent disciplines. Moreover, the articles that follow do attempt, often in bold and innovative ways, to make some inroads on these questions. They explore aspects of these issues from a number of vantage points. From philosophy, we see discussions of what explanations are and how they contrast and relate across different established sciences, as well as other domains (Thagard; Clark; Wilson & Keil). Philosophers can also provide a more general, metaphysical perspective on the kinds of structures and causal patterns that fit with epistemic views of how we can possibly grasp those structures (Glymour). From a more computational perspective we see discussions as to how notions of explanation and cause can be instantiated in a range of possible learning and knowledge systems, and how they can be connected to the causal structure of the world (Simon; Glymour; Thagard). Finally, from psychology we see discussions of how adults mentally represent, modify and use explanations; how children come to acquire them; and what sorts of information, if any, humans are naturally predisposed to use in building and discovering explanations (Gopnik and Brewer, Chinn and Samarapungavan; Wilson and Keil). More importantly, however, all of these articles show the powerful need to cross traditional disciplinary boundaries in order to be able to develop satisfactory accounts of explanation. Every article draws on work across several disciplines, and in doing so, develops insights not otherwise possible.

Simon asks how it is that we can discover explanations, an activity at the heart of science, and move beyond accounts that merely describe events to those that explain their structure. He applies his “physical symbol system hypothesis” in his answer, considering classes of information-processing mechanisms that might discover explanations, and how computational models might inform psychological ones. He also considers patterns in the history and philosophy of science and their relations to structural patterns in the world, such as the phenomenon of nearly-decomposable systems and their more formal properties. These issues in turn bring in questions concerning the social distribution and sharing of knowledge.

Glymour focuses on the question of how we learn about causal patterns, a critical component in the emergence of most explanations. Building on developments in computer science that consider conditional probability relations in multi-layered causal networks, he considers how a combination of tabulations of probability

information and a more active interpretative component allow the construction of causal inferences. This discussion raises the natural question of how humans might operate on such multi-layered causal networks; an area largely unexplored in experimental research. Glymour turns to work by Cheng on covariation judgments to build links between computational and psychological approaches, and to set up a framework for future experiments in psychology.

Thagard focuses on the related and long-standing problem of how one makes the inference from correlation to causation. He suggests that some sense of mechanism is critical to make such inferences and discusses how certain causal networks can represent such mechanisms and thereby license the inference. His discussion covers psychological work on induction, examines epidemiological approaches to disease causation, explores historical and philosophical analyses of the relations between cause and mechanism, and considers computational problems of inducing over causal networks.

Clark discusses how many phenomena in biology and cognitive science seem to arise from a complex, interconnected network of causal relations that defy simple hierarchical or serial characterizations, and which are often connected in recurrent interactive loops with other phenomena. He responds to arguments that such patterns imply that one should reject any notions involving internal causal factors, such as genes or mental representations. Clark argues that this is an excessively strong reaction to these complex systems, and he shows how models in cognitive science and biology do not need to reject such explanatory schema. His discussion thereby links questions about the philosophy of science to the practice of cognitive science.

Gopnik addresses the phenomenology of what she calls the “theory formation system”, developing an analogy to biological systems that seem to embody both drives and a distinctive phenomenology. In discussing this phenomenology, her account blends together psychological and philosophical issues and illustrates how developmental and learning considerations can be addressed by crossing continuously between these two disciplines. Her paper also brings in considerations of the evolutionary value of explanation, and why it might be best conceived of as a drive similar in many respects to the more familiar physiological drives associated with nutrition, hydration and sex.

Brewer, Chinn and Samarapungavan ask how explanations might be represented and acquired in children, and how they compare to those in scientists. They propose a general framework of attributes for explanations, attributes that would seem to be the cornerstones of good explanations in science, but which perhaps surprisingly also appear to be the cornerstones of explanation in even quite young children. There therefore seem to be natural constraints on explanations that can be found both in history and philosophy of science and in its earliest origins in development. At the same time, explanations in science are different from both those in every day life and from those in the minds of young children, and Brewer et al. also discuss what may differ and why.

Finally, our own article attempts to more fully characterize what explanations are and how they might differ from other ways in which we can partially grasp the causal structure of the world. We suggest that traditional discussions of explanation in the philosophy of science introduce conceptions of explanation that give us mere “shadows” of explanation when we consider explanation in everyday life, and that one of explanation’s surprising features is its relative psychological “shallowness”. Based on psychological work, we suggest that most common explanations, and probably far more of hands-on science than one might suspect, have a structure that is more implicit and schematic in nature than is suggested by more traditional accounts. We argue that this schematic and implicit nature is fundamental to explanations of value in most real world situations, and show how this view is compatible with our ability to tap into causal structures in the world and to engage in explanatory successes.

What is striking in all of these articles is the inability to pigeonhole them as coming from the viewpoint of just one discipline. It seems clear that the task of explaining explanation reveals how it is indeed a problem of cognitive science, and how it is especially appropriate for a special issue of a journal like *Minds and Machines*.

FRANK KEIL AND ROB WILSON

## References

- Carey, S. (1985). *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.
- Carey, S., Johnson, S. C. and Levine, K. (1993), *Conceptual Structure of Adults/Adolescents with Williams Syndrome*. Paper presented at the Society for Research in Child Development, New Orleans.
- Chen P. (1997), “From Covariation to Causation: A Causal Power Theory”, *Psychological Review*.
- Udwin, O. and Yule, W. (1991). A Cognitive and Behavioral Phenotype in Williams Syndrome, *Journal of Clinical and Experimental Neuropsychology* 13, 232–244.