Oxford Handbooks Online

Explanation and Abductive Inference

Tania Lombrozo

The Oxford Handbook of Thinking and Reasoning
Edited by Keith J. Holyoak and Robert G. Morrison

Print Publication Date: Mar 2012
Subject: Psychology, Cognitive Psychology, Cognitive Neuroscience
Online Publication Date: Nov 2012 DOI: 10.1093/oxfordhb/9780199734689.013.0014

Abstract and Keywords

Everyday cognition reveals a sophisticated capacity to seek, generate, and evaluate explanations for the social and physical worlds around us. Why are we so driven to explain, and what accounts for our systematic explanatory preferences? This chapter reviews evidence from cognitive psychology and cognitive development concerning the structure and function of explanations, with a focus on the role of explanations in learning and inference. The findings highlight the value of understanding explanation and abductive inference both as phenomena in their own right and for the insights they provide concerning foundational aspects of human cognition, such as representation, learning, and inference.

Keywords: explanation, understanding, abduction, abductive inference, inference to the best explanation, self-explanation

Introduction

This chapter considers a ubiquitous but understudied aspect of human cognition: explanation. Both children and adults seek explanations constantly. We wonder why events unfold in particular ways, why objects have specific properties, and why others behave as they do. Moreover, we have strong intuitions about what requires explanation and what counts as an adequate answer. Where do these intuitions come from? Why are we so driven to explain the social and physical worlds around us? And more generally, what is the role of explanation in cognition?

While there is no consensus on how to address these questions, research in cognitive development and cognitive psychology is beginning to make headway in finding answers (Keil, 2006 Keil & Wilson, 2000; Lombrozo, 2006; Wellman, 2011). These new developments complement large but focused literatures in social psychology (e.g., Anderson, Krull, & Weiner, 1996; Heider, 1958; Malle, 2004), education (e.g., Chi, 2000; Roscoe &Chi, 2008), and philosophy (e.g., Salmon, 1989; Woodward, 2010), which have respectively emphasized explanations for human behavior, explanations in pedagogy, and explanations in science. Explanation has also been of interest in other cognitive science disciplines, such as artificial intelligence, where the generation of explanations has been proposed as a mechanism for learning (e.g., DeJong, 2004).

In this chapter, I focus on recent research in cognitive psychology and cognitive development. I begin by motivating the study of explanation, what explanations are, and their importance for understanding cognition. This is followed by more extensive discussions of the roles of explanation in learning and inference. While explanations and their effects are quite diverse, the findings support a common picture according to which the process of explaining recruits prior beliefs and a host of explanatory preferences, such as unification and simplicity, that jointly constrain subsequent processing. I conclude (p. 261) by considering the current status of research on explanation and highlighting promising directions for future research.

Why Explanation Matters for Cognition

The study of explanation is motivated both by everyday experience and by theoretical considerations. Seeking, generating, and evaluating explanations is a central human preoccupation from an early age, and anecdotal evidence suggests that explanations play an important role in motivating discovery, communicating beliefs, and furthering understanding. Much of the recent interest in explanation, however, stems from theoretical and empirical developments that place explanation at the core of claims about conceptual representation, learning, and inference.

Since at least the 1980s, a prominent strand of thinking within cognitive science has postulated structured mental representations that share many of the properties of scientific theories (e.g., Carey, 1985; Gopnik & Meltzoff, 1997; Murphy & Medin, 1985; Wellman & Gelman, 1992; see Gelman & Frazier, Chapter 26). These so-called naïve, folk, or intuitive theories are semicoherent bodies of belief that support prediction, explanation, and intervention for a specified domain of application, such as physics or psychology. Almost all advocates for this approach have appealed to explanation in defining intuitive theories. For example, theories have been identified with “any of a host of mental explanations” (Murphy & Medin, 1985, p. 290) and characterized in terms of “laws and other explanatory mechanisms” (Carey, 1985, p. 201). It is clear, then, that an

adequate account of intuitive theories rests on an adequate account of explanation (see also Lombrozo, 2010a; Machery, 2009; Margolis, 1995).

In parallel with a focus on intuitive theories, accounts of concepts and categorization have increasingly emphasized "explanation-based" reasoning in contrast to purely statistical or similarity-based reasoning (see Rips et al., Chapter 11). For example, categorization and category learning have been described as "special cases of inference to the best explanation" (Rips, 1989, p. 53), with a concept invoked "when it has a sufficient explanatory relation to an object" (Murphy & Medin, 1985, p. 295). And in tasks such as property generalization, deviations from patterns of reasoning based on similarity are often explained by appeal to causal or explanatory beliefs (e.g., Murphy & Allopenna, 1994; Rehder, 2006), pointing to the need for an account of what makes some beliefs "explanatory," and how those beliefs influence reasoning.

Finally, there is increasing evidence that prior beliefs significantly impact how reasoners learn and draw inferences. Generating and evaluating explanations may be mechanisms by which prior beliefs are brought to bear on a given task (Lombrozo, 2006), especially in knowledge-rich domains. The study of explanation is therefore not only of value in its own right, but as a window onto foundational aspects of cognition ranging from conceptual representation to learning and inference.

Defining Explanation

A natural place to begin the study of explanation is with a clear characterization of the object of study. In other words, what *is* explanation? Unfortunately, this innocuous question has proven quite complex. Some researchers define explanations as answers to why- or how-questions (e.g., Wellman, 2011), others as judgments about why an outcome occurred (e.g., Krull & Anderson, 2001), yet others as hypotheses positing causes that have what is being explained as their effects (e.g., Thagard, 2000; see also Einhorn & Hogarth, 1986). In all likelihood, explanation is not unitary: Research on explanation almost certainly spans multiple kinds of judgments and distinct cognitive mechanisms. "Explanation" is likely to be a family-resemblance term, picking out a cluster of related phenomena.

A first step toward precision, if not definition, is to distinguish explanation as a *product* from explanation as a *process* (see also Chin-Parker & Bradner, 2010). As a product, an explanation is a proposition or judgment, typically linguistic, that addresses an explicit or implicit request for an explanation. As a process, explanation is a cognitive activity that aims to generate one or more explanation "products" but need not succeed in order to be engaged. Most theories of explanation from philosophy are about explanations as products, whereas empirical research has been more mixed, with some research focusing

on the characteristics of the product, and other research on the characteristics and consequences of the process.

A second useful distinction is between a “complete” explanation and a “selected” explanation (for related distinctions see Goodman et al., 2006; Railton, 1981). To illustrate, consider an explanation for Carl's decision to order a slice of carrot cake over chocolate cake. In most contexts, noting a stable preference (“carrot cake is Carl's favorite”) can (p. 262) be sufficient to explain the decision, but this explanation rests on a variety of background assumptions (for example, that a preference for carrot cake would lead someone to choose carrot cake over chocolate cake), as well as a much deeper causal structure that could potentially be traced back to the origins of the universe. A selected explanation will correspond to the small subset of this explanatory structure that is identified as “the” explanation in a given context: probably the cake preference, but probably not the big bang.

This distinction is important in understanding explanation as both product and process. When considering explanation as a product, there is the selected explanation itself—what an individual might offer in response to a why-question—as well as a more complete explanation, which includes the cognitive infrastructure that supports the selected explanation. Of course, when it comes to human cognition, the complete explanation could be quite incomplete—it's likely that human explanatory practice proceeds in the absence of fully specified explanations, even if there is always a more complete explanation that underlies the selected explanation (Keil, 2003; Rozenblit & Keil, 2002). When considering explanation as a process, it is useful to distinguish two distinct (although not necessarily independent) inferential steps: one to the complete explanation, and a second “selection process” to identify the selected explanation (which could, in principle, be equivalent to the complete explanation).

Potential insight into how selection occurs comes from the observation that explanations are typically contrastive: They account for one state of affairs in contrast to another (see e.g., Chin-Parker & Bradner, 2010; Garfinkel, 1990; Hart & Honore, 1985; Hilton, 1996; Hilton & Slugoski, 1986; Mackie, 1980; van Fraassen, 1980). The role of an implicit contrast is illustrated by an explanation attributed to Willie Sutton, a notorious bank robber. When asked why he robbed banks, he explained: “Because that's where the money is.” The response is humorous in part because it addresses an inappropriate contrast: Sutton has explained why he robs banks *rather than robbing other places*, but he has failed to adequately answer the question most likely intended, namely why he robs banks *rather than not robbing banks*. The contrast provides a constraint on what should figure in a selected explanation: Explanations typically identify conditions that differentiate what is being explained (the explanans) from a counterfactual case in which the contrast holds. Aspects common to the explanans and to the contrast likely figure in the complete explanation, which supports the selected explanation but need not be found explanatory. For example, Carl's choice of carrot cake over chocolate cake is poorly explained (if at all) by noting that “cake is delicious,” as this piece of background does

not explain why he chose carrot cake *as opposed to chocolate cake*. Recognizing the contrastive nature of explanation also reveals that a full specification of what is being explained involves identifying assumptions that are implicit in an explanation request.

The distinctions between explanation as product and process and complete versus selected explanations highlight one of the challenges in the study of explanation: providing cohesive theories that are nonetheless sensitive to this diversity. Explanatory products and processes are clearly related and mutually constraining, as are complete and selected explanations. Yet a failure to recognize these distinctions can make it difficult to relate different strands of research or to appreciate their implications for cognition.

The Structure of Explanations

Within philosophy, there have been several systematic attempts to provide definitions or precise specifications of what counts as an explanation (the product), and in particular the relationship that must obtain between an explanation and what it explains. Although typically intended as normative claims about explanation in science, these theories provide a useful starting point for psychological investigation and help identify the space of possible approaches. Initiating the contemporary study of explanation in philosophy of science, Hempel and Oppenheim, (1948) proposed the deductive-nomological (DN) account of explanation, according to which an explanation involves initial conditions and general laws from which one can deduce what is being explained (the explanandum). This account was later extended to address inductive rather than deductive cases (the inductive-statistical [IS] account; see Hempel, 1965), but it still faced serious problems that led most philosophers to alternative accounts (see Salmon, 1989, for a review).

Currently, there are three popular approaches to explanation within philosophy of science. The two most prominent are causal theories (see Buehner & Cheng, Chapter 12), according to which explanations identify the cause(s) for an event or property (e.g., Salmon 1984; Woodward, 2003), and (p. 263) subsumption or unification theories, which claim that an explanation subsumes the explanandum under an explanatory pattern, which can but need not be a law or counterfactual-supporting generalization (e.g., Friedman, 1974; Kitcher, 1989). These two approaches have also been fruitfully combined. For example, Strevens, (2008) proposes a causal theory of explanation but relies on criteria involving unification to solve the selection problem of identifying a selected explanation from a complete one. A third approach, related to causal approaches, is to identify explanations with a description of causal *mechanisms*, where a mechanism consists of components, operations, and their (often hierarchical) organization, which realize what is being explained (e.g., Bechtel, 2008; Craver, 2007; Darden, 2006).

While many everyday and scientific explanations in fact satisfy all of these theories, the approaches differ in their account of *why* the explanation is explanatory. For example, consider explaining a disease by appeal to a virus. According to a causal theory, an occurrence of the disease can be explained by appeal to the presence of the virus because the virus causally accounts for the occurrence of the disease. According to a subsumption/unification theory, the virus explains the disease by subsuming the instance being explained under a broader generalization—that the particular virus leads to that particular disease—which in turn conforms to a broader generalization about viruses causing diseases, and so on to potentially greater levels of abstraction. While multiple theories can thus accommodate a wide range of common cases, they face distinct challenges. In general, causal and causal mechanism approaches have a difficult time handling cases of explanation that do not seem to involve causation at all, such as mathematical explanations. In contrast, subsumption/unification theories face the problem of specifying which kinds of patterns or unifications count as explanatory. A strategy potentially amenable to both causal and subsumption theories is to identify some kind of higher order, asymmetric dependence relationship or generative process that can subsume both causal and mathematical or formal relationships, and serve as a foundation for a theory of explanation with broader scope.

Causal, subsumption/unification, and mechanism-based approaches all find some empirical support within psychology. Many investigations of explanation have been restricted to causal cases, but some accounts of the effects of explanation on learning appeal to subsumption and unification (see later section on “Explanation and Learning”). There is also evidence that knowledge of mechanisms is often invoked in explanations and constrains inference (e.g., Ahn & Bailenson, 1996; Ahn, Kalish, Medin, & Gelman, 1995; Koslowski, 1996; Shultz, 1980). However, no studies (to my knowledge) have attempted to differentiate these theories empirically by isolating cases for which the theories generate different predictions.

The Functions of Explanation

Intertwined with questions about the structure of explanations are those about its functions. Why are people so motivated to explain, and what accounts for our systematic explanatory preferences? While many plausible functions for explanation have been proposed, both philosophers and psychologists have emphasized that explanations could be valuable because they scaffold the kind of learning that supports adaptive behavior. For example, Craik, 1943 described explanation as “a kind of distance-receptor in time, which enables organisms to adapt themselves to situations that are about to arise.” Heider (1958) suggested that we explain events in order to relate them to more general processes, allowing us “to attain a stable environment and have the possibility of

controlling it.” In other words, explanations put us in a better position to predict and control the future.

Gopnik, (2000) provocatively compares explanation to orgasm, suggesting that the phenomenological satisfaction of explanation is our evolutionarily provided incentive to engage in theory formation, as orgasm is to reproduction. She thus explains a characteristic of the process of explanation—its typical phenomenological outcome—by appeal to the function of its products—namely useful theories about the world. But the very process of engaging in explanation could have additional benefits. For example, attempting but failing to produce accurate explanations could nonetheless support future learning (e.g., Needham & Begg, 1991) and guide future inquiry (e.g., Legare, in press), and the effects of explanation could differ depending on whether the explanations are self-generated (involving process plus product) or provided (involving only product; see, e.g., Brown & Kane, 1988; Crowley-Siegler, 1999; Rittle-Johnson, 2006 for relevant discussion and findings).

The hypothesis that explanation supports adaptive behavior predicts that seeking, generating, and evaluating explanations should have systematic cognitive **(p. 264)** consequences. The rest of the chapter is principally devoted to documenting these consequences. Before moving on, however, a few caveats are in order. First, explanations need not share a single, common function. Accordingly, the section that follows considers different kinds of explanations and their respective consequences for cognition. And while the present focus is on the roles of explanation in learning and inference, it is worth acknowledging that explanation certainly serves additional functions, as in persuasion and the assignment of blame.

A second caveat is that whatever the functions of explanation, it is unlikely that explainers have the explicit goal of fulfilling them. The basis for explanatory judgments could be opaque to introspection or based on indirect cues, no matter that their ultimate function is to achieve particular goals. And finally, it is quite likely that many of the properties of explanation stem not from any particular benefit (be it the result of natural selection or learning) but rather arise as a side effect of other cognitive characteristics. For example, a preference for simpler explanations could be a side effect of limited working memory. These caveats, however, do not detract from the value of examining the role of explanation in learning and inference.

Different Kinds of Explanations

Not all explanations will necessarily have a common structure or function. In fact, there are a variety of proposals suggesting fundamentally different kinds of explanations. For example, within philosophy, explanations for particular events or properties are often distinguished from explanations for laws or regularities, as are “how possibly” explanations—which explain how something *could* have come about—from “how actually”

explanations—which explain how something *in fact* did come about (e.g., Salmon, 1989). Within psychology, advocates for domain specificity have suggested that different domains involve unique explanatory schemata (e.g., Carey, 1985; Wellman & Gelman, 1992). And explanations for why something is the case can be distinguished from explanations for why someone believes something to be the case (e.g., Kuhn, 2001). Depending on the particular account of explanation in question, these distinctions can correspond to explanations that differ in structure, in function, in both, or in neither.

A taxonomy that has proven particularly fruitful in psychology, with roots in Aristotle (see also Dennett, 1987), identifies (at least) three kinds of explanations: mechanistic explanations, which explain by appeal to parts and processes; teleological or functional explanations, which cite functions or goals; and formal explanations, which cite kind or category membership. For example, consider explanations for why a particular tire is round. Explaining the roundness by appeal to the tire's manufacturing process would qualify as mechanistic; explaining by appeal to its function in generating efficient movement would qualify as functional; and explaining by appeal to the objects' category membership (“it's round because it's a tire”) would qualify as formal (see Lombrozo & Carey, 2006, for an extended discussion of mechanistic and functional explanations; see Prasada & Dillingham, 2006, for a discussion of formal explanations).

A growing body of evidence suggests that these three kinds of explanation correspond to distinct cognitive processes and representations. Keil (1992, 1994, 1995)for example, posits distinct explanatory stances or “modes of construal” corresponding to mechanistic and functional explanations, and finds that children adopt explanatory stances preferentially in different domains. Developing earlier suggestions by Sully (1900 and Piaget (1929) , Kelemen (1999 suggests that functional explanations are an explanatory “default” that comes more naturally to children than do mechanistic explanations. It appears that children in fact do prefer such explanations “promiscuously,” as do adults who have not been exposed to alternative scientific explanations (Casler & Kelemen, 2008), who are required to respond under speeded conditions (Kelemen & Rosset, 2009), or who suffer from Alzheimer's disease (Lombrozo, Kelemen, & Zaitchik, 2007).

Although functional explanations appeal to effects rather than to causal mechanisms, they are not necessarily inconsistent with causal theories of explanations. In fact, Lombrozo and Carey, 2006Carey find evidence that functional explanations are only accepted when the function invoked in the explanation played a causal role in bringing about what is being explained (or more precisely, when the function is a token of a type that played a causal role in bringing about what is being explained; see also Wright, 1976, for analysis from philosophy). Thus, a tiger's stripes can be explained by appeal to camouflage because the appeal to camouflage is shorthand for a complex causal process, whereby the stripes of past tigers supported camouflage, and this in turn played a causal role in the existence of the current, striped tiger. **(p. 265)**

Mechanistic and functional explanations can thus both be construed as causal explanations, but there is reason to think they reflect differences in underlying reasoning. First, while functional explanations ground out in a causal story, that story can be unboundedly complex and largely unknown. Most people are happy to explain a tiger's stripes by appeal to camouflage without any knowledge of tigers' evolutionary history, and with almost no knowledge of how natural selection operates (e.g., Shtulman, (2006). This suggests that information about causal mechanisms plays a different role for functional explanations than for mechanistic explanations. Second, mechanistic and functional explanations highlight different information. Mechanistic explanations privilege causal mechanisms and underlying constituents, while functional explanations privilege functions, intentions, goals, and design.

Consistent with this observation, Lombrozo, (2010b) finds that the criteria for causal ascription differ depending on whether a causal system is construed functionally or mechanistically: When a system is construed mechanistically, causal ascriptions are more sensitive to the mechanism of transmission by which a cause brought about an effect. For example, a causal factor that contributed to an outcome indirectly, by preventing another factor from preventing the outcome, will receive a lower causal rating from a "mechanistic stance," but not necessarily from a "functional stance." Additional findings suggest that functional and mechanistic explanations have differential effects on categorization (Lombrozo, 2009) and generalization (Lombrozo & Gwynne, unpublished data), and are discounted asymmetrically (Heussen, 2010).

The unique explanatory contributions of mechanistic and functional explanations may be familiar to psychologists acquainted with Marr's levels of analysis (Marr, 1982). Marr proposed that psychological phenomena can be explained at three levels, including the *computational*, which involves a specification of the goals of a system; the *algorithmic*, which identifies the representations and processes involved in carrying out the specified goal; and the *implementational*, which characterizes how the representations and processes are physically instantiated. In specifying the goals of a cognitive system, a computational-level analysis typically supports functional explanations, while the implementation level analysis typically supports mechanistic explanations. The algorithmic level can support both kinds of explanations. For example, representations of binocular disparity can be explained (functionally) by appeal to their role in computing depth information, and also (mechanistically) by appeal to optical, retinal, and cortical processes.

The final category of "Aristotelian" explanation that has received empirical attention is formal explanation, the study of which has been pioneered by Prasada and colleagues (Prasada & Dillingham, (20062009). They find that only some properties can be explained by appeal to category membership. For example, participants are willing to explain a tire's roundness by appeal to its category membership ("it's round because it's a tire"), but much less willing to thus explain a tire's blackness ("it's black because it's a tire"), even though both share the same statistical association with tires. Properties that support

formal explanations have a variety of related characteristics, such as licensing normative claims (e.g., “tires ought to be round”).

In sum, explanations can be classified in a number of ways. The distinction between mechanistic, functional, and formal explanations has proven particularly fruitful, as each type of explanation appears to highlight different information and have unique cognitive consequences. This distinction could also underlie differences in the types of explanations typically observed across domains (for discussion, see Lombrozo & Carey, 2006). However, there are likely to be a number of alternative classifications with potential implications for our understanding of both the structure and functions of explanations. For example, the discussion so far has focused on how *selected* explanations—the *products*—can be productively classified, but complete explanations and explanatory processes could be similarly varied. Documenting and understanding the heterogeneity of explanation is thus an important focus for future research.

Explanation and Learning

Given the intimate relationship between explanation and understanding, it is no surprise that explanation has a profound impact in learning. There are at least three ways in which explanation can influence learning. First, there is the matter of which explanations are sought, which constrains what one learns about the environment. For example, upon first encountering an elephant you're likely to wonder why it has a trunk, but less likely to wonder why the number of its legs is a perfect square. Second, processes involved in the evaluation of explanations can influence what is learned (p. 266) from provided explanations, be it in educational or everyday situations. And third, the very process of generating explanations, be it for oneself or others, can influence one's own understanding and ability to generalize to novel contexts. These three impacts on learning are considered in turn.

Although explanation seeking is pervasive, people are highly selective in what they seek explanations about. This selectivity is driven in large part by what the learner already knows. For example, it seems unlikely that someone will seek an explanation for why fire trucks are red if that person already knows (or believes she knows) the answer. This is not surprising if explanation seeking is a mechanism for learning. However, generating questions requires knowing enough to appreciate what one doesn't know (Miyake & Norman, 1979). Questions are thus highly constrained by prior beliefs. For example, we ask questions for which we expect sensible answers—even young children are more likely to ask what artifacts “are for” than what animals “are for” (Greif, Kemler-Nelson, Keil, & Guiterrez, 2006).

One attractive hypothesis is that people ask questions about events or properties that violate their expectations (Isaacs, 1930; Sully, 1900; see alsoBerlyne, 1954; Chouinard, 2007) or concern abnormal events Isaacs, 1990; Hilton & Slugoski, 1986; see alsoBerlyne,

1954)—a natural prediction if explanation is to guide learners beyond what they already know. Consistent with this proposal, analyses of why-questions reveal a high prevalence of negation (e.g., “Why *didn't* X occur?”), suggesting that the questioner is grappling with an unexpected observation (Hood & Bloom, 1979;). And in experimental contexts, children are more likelbut see Callanan & Oakes, 1992y to explain why a block did or did not make a machine light up when the observed outcome is inconsistent with previous training (Legare, Gelman, & Wellman, 2009). However, it is quite difficult to specify when an event counts as unexpected or surprising—it surely *isn't* the case that observations with low prior probability always prompt explanation, as most observations are in fact highly unlikely (consider, for example, any long sequence of coin flips: heads, heads, tails, heads, tails …).

Questions are also, of course, guided by interests: Children seem most likely to ask about the social world (Hickling & Wellman, 2001; Hood & Bloom, 1979;), although questions about the physical world are not uncommon (Callanan & Oakes, 1992). There is also some evidence that from preschool to high school, children become increasingly likely to ask questions motivated by potential applications rather than by general curiosity (Baram-Tsabari & Yarden, 2005). Finally, questions are guided by the goals of the learner. Students aiming to learn about musical instruments in order to design a novel instrument, for example, ask different questions from those exposed to the same material with the goal of identifying a promising musical group (Graesser, Langston, & Baggett, 1993). Providing a general characterization of what, when, and why people ask is complicated by the fact that interests and question-asking are highly contingent on both the explanation seeker and the context.

Once an explanation has been offered, what are the consequences for the recipient? First, explanations are often accompanied by a sense of understanding. For example, both undergraduates and clinicians who could explain an individual's symptoms perceived those symptoms to be more “normal” (Ahn, Novick, & Kim, 2003). And in some cases, provided explanations can be highly beneficial. As one example, children's performance on false belief tasks is related to how often their mothers explain mental states (Peterson & Slaughter, 2003). Unfortunately, however, the *sense* of understanding fostered by explanations is often an unreliable guide to actual understanding (Trout, (2002, 2008), and merely receiving an explanation is often insufficient to generate actual understanding. In particular, explanations provided in instructional contexts are frequently ineffective (Wittwer & Renkl, 2008). Explanations are more likely to support learning if they are appropriately tailored to what the learner already knows and appeal to general concepts or principles (see Wittwer & Renkl, 2008, for review). Explanations can also be more effective when they occur in the context of an interactive cognitive activity (Chi, 2009), in part because the recipient is able to request additional information and receive feedback. For example, young children are more likely to repeat questions after receiving a nonexplanation or an explanation they deem inadequate (Frazier, Gelman, & Wellman,2009; see also Chouinard, 2007), and adult learners will request

further elaboration when experts provide explanations that underestimate their knowledge (Wittwer, Nuckles, & Renkl, 2008).

Perhaps surprisingly, *generating* explanations can be a more effective mechanism for learning than *receiving* explanation. This phenomenon has been demonstrated in the context of peer tutoring, where tutors often profit more than tutees (e.g., Hooper, 1992; Roscoe & Chi, 2008; 1995Ross & Cousins, 1995;). The learning benefit of engaging in explanation—be it to oneself or to others—is known as the self-explanation effect (Chi, Bassok, Lewis, Reimann, & Glaser, 1989) Chi, de Leeuw, Chiu, & LaVancher,1994) and has been found for preschoolers through adults, for a range of educational materials, and for both declarative and procedural knowledge (for review, see Fonseca & Chi, 2010). In a typical experiment, one group of participants is prompted to explain to themselves as they study an expository text or worked examples, such as math problems. These participants are compared with those in one or more control groups who study the same material without a prompt to explain, often with an alternative task (e.g., thinking aloud) or matched for study time. The typical finding is that participants who explain outperform their nonexplaining peers on a posttest, with the greatest benefit for transfer problems that require going beyond the material presented.

Effects of self-explanation have also been demonstrated in complex domains that arguably require a conceptual change in development. For example, prompting children to explain the correct response on Piagetian number conservation tasks leads to greater improvement than prompting them to explain their own (potentially incorrect) response, or than merely receiving feedback on the correct response (Siegler, 1995). Similarly, prompting children to explain why a character searched for an item in a particular location in a false belief task leads to greater improvement than merely predicting the character's behavior and receiving feedback, with benefits that extend to other theory of mind tasks (Amsterlaw & Wellman, 2006). These findings attest to the power of explanation as an engine for learning.

A variety of plausible mechanisms underlying the effects of explanations on learning have been proposed, many of which could operate in concert. For example, generating explanations can help learners translate declarative knowledge into usable procedures (Chi et al., 19891994), integrate new material with prior beliefs (Ahn, Brewer, & Mooney, 1992; Chi et al., 1994; Lombrozo, 2006), identify and repair gaps in knowledge (Chi, 2000), or resolve potential inconsistencies (Johnson-Laird, Girotto, & Legrenzi, 2004). Other researchers have suggested that prompts to explain can increase efforts to seek explanations, leading to deeper processing, greater engagement with the task, and the reinforcement of successful strategies over unsuccessful alternatives (Siegler, 2002). More recent work in cognitive development suggests that explanations can sometimes scaffold causal learning by presenting children with an easier task than prediction, as explaining is supported by knowledge of the outcome or property being explained (Wellman & Liu, 2007). As evidence, successful explanations precede successful predictions by preschoolers reasoning about an agent's choice between a contaminated

and an uncontaminated food (Legare, Wellman, & Gelman, 2009; see also Amsterlaw & Wellman, 2006).

In comparing participants prompted to explain with those who are not so prompted, much of the research on explanation and learning focuses on explanation as a process rather than a product. This approach sidesteps questions about what kinds of utterances count as explanations. Research that has coded participant utterances for the quantity and content of explanations has tended to adopt a very broad conception of explanation, for example, including most articulated inferences that go beyond the material provided (Chi et al., 1989, 1994). It can thus be difficult to relate aspects of the structure of explanations, such as those posited by theories of explanation from philosophy, to the functional consequences of explanation for learning.

A few more recent strands of research, however, aim to relate structural accounts of explanation to functional consequences, and in so doing to elucidate why explanation has the particular learning benefits observed. In particular, research has focused on two widely accepted features of explanations: explanations tend to relate the particular property or event being explained to broader principles or regularities, and explanations often invoke causal relationships and causal mechanisms (see Woodward, 2010 for review from philosophy; see Lombrozo, 2006 for review from psychology). These properties correspond to important features of subsumption/unification and causal theories of explanation from philosophy, respectively.

Illustrating the role of subsumption and unification, Williams and Lombrozo (2010) propose the subsumptive constraints account, according to which explanation can facilitate learning by encouraging learners to understand what they are trying to explain in terms of broad, unifying generalizations (see also Wellman & Liu, 2007). If explaining exerts this particular constraint on the kind of structure learners seek, engaging in explanation should drive learners to discover and explicitly represent such **(p. 268)** generalizations. Williams and Lombrozo (2010)tested this prediction in several experiments involving category learning. Compared to control conditions in which participants described category members, thought aloud during study, or engaged in free study, participants who explained were more likely to discover subtle, broad regularities underlying the category structure. Moreover, because the subsumptive constraints account predicts a selective advantage for explanation in discovering patterns, and not an "all-purpose" benefit in learning, it makes the novel prediction that explaining can *impair* learning when the material being explained does not support explanatory patterns, and recent findings confirm this prediction (Williams, Lombrozo, & Rehder, 2010 Kuhn & Katz, 2009;; see also Dawes, 1999).

Illustrating the role of causation and causal mechanisms in mediating the role of explanation in learning are recent findings from cognitive development. Legare, Wellman, and Gelman (2010) found that children's explanations posited unobserved causal mechanisms, such as germs, that were not spontaneously considered as a basis for prediction. This result suggests that explanations direct children to aspects of causal

structure that might not be taken into account in the absence of explanatory processes (see also Legare, Gelman, & Wellman, 2009). Children who are prompted to explain how a novel, mechanical toy works also outperform peers who are prompted to observe the toy for a matched amount of time on measures that assess an understanding of causal mechanisms, but not on measures that involve memory for mechanism-irrelevant details, such as the toy's colors (Legare & Lombrozo, unpublished data). Again, the findings suggest that explanation is not an all-purpose strategy for learning, but rather a highly selective process that influences precisely what a learner discovers and represents.

A great deal remains to be learned about the mechanisms by which explanation influences learning and how distinct mechanisms interact. Characterizing these mechanisms is important not only for understanding explanation but also for understanding the very nature of human learning and representation. One reason the self-explanation effect is so intriguing is because it challenges a simple picture of learning, modeled after learning from observations or testimony, according to which learning is identified with the acquisition of new information from the external world. Learning by explaining to oneself or to others—much like thought experiments—reveals a kind of learning that involves the reorganization, rerepresentation, or inferential consequences of what is already known. Understanding this kind of learning could well provide a better template for understanding all types of learning (for relevant discussions of learning see Rips et al., Chapter 11; Buehner & Cheng, Chapter 12; Holyoak, Chapter 13; Bassok & Novick, Chapter 21; and Koedinger & Roll, Chapter 40).

Explanation and Inference

Although learning and inference are likely to involve similar mechanisms, research on explanation and inference has proceeded largely independently of research on explanation and learning. The former strand of research is motivated by the observation that everyday inferences face the problem of underdetermination: Many hypotheses are possible given what we know about the world. We generally don't know for certain why the economy rallied, why the bus was late, whether a colleague's excuse for a missed deadline was fact or fabrication, or whether Mrs. Peacock or Mr. Green committed the murder. Yet many decisions rest on identifying the best hypothesis or the probability of a given hypothesis. Explanation has thus been proposed as a mechanism for guiding such judgments, often referred to as “abductive inferences.”

The term “abduction” is associated with Charles Sanders Peirce, who distinguished abductive inferences, which posit explanatory hypotheses, from the broader class of inductive inferences, which also include inferences from a sample to a population, such as inferring that all swans are white after observing many white swans (Burch, 2010; see also Magnani, 2001). Subsequent scholars have not reliably distinguished abduction from induction, instead focusing on inferences to particular explanations. For example, Harman (1965) introduced the notion of “inference to the best explanation,” according to which “one infers, from the fact that a certain hypothesis would explain the evidence, to the truth of that hypothesis … one infers, from the premise that a given hypothesis would provide a ‘better’ explanation for the evidence than would any other hypothesis, to the conclusion that the given hypothesis is true” (p. 324). In other words, one relies on the existence or quality of an explanation to guide inference, an idea that has since been pursued both theoretically and empirically (see Lipton, 2004, for a nice treatment in philosophy).

Within psychology, there is mounting evidence that the quality of an explanation does serve as (p. 269) a guide to its probability. Factors that boost an explanation's perceived quality include simplicity (Bonawitz & Lombrozo, in press; Lagnado, 1994; Lombrozo, 2007; Read & Marcus-Newhall, 1993; Thagard, 1989), breadth (Preston & Epley, 2005; Read & Marcus-Newhall, 1993; Thagard, 1989), consistency with prior knowledge (Thagard, 1989), and coherence (Pennington & Hastie, 1993). Several of these factors have also been shown to influence an explanation's perceived probability. For example, Lombrozo (x2007) presented adults with novel diseases and symptoms. Participants learned about a patient with two symptoms that could be explained by appeal to a single disease (the “simple” explanation) or by appeal to the conjunction of two diseases (the “complex” explanation). In some conditions, participants were also presented with the base rates for the diseases, either in the form of summary frequencies or experienced cases. There were several notable results: *(a)* in the absence of base-rate information, participants overwhelmingly preferred the simple explanation; *(b)* when base-rate information was provided, the complex explanation had to be quite a bit more probable

than the simpler explanation for a majority of participants to prefer it; *(c)* participants who selected the simple explanation when it was unlikely to be true overestimated the base-rate of the disease figuring in the simple explanation; yet *(d)* when the complex explanation was explicitly stated to be more likely, participants overwhelmingly chose it. These results suggest that in the face of probabilistic uncertainty, people treat simplicity as a probabilistic cue commensurate with base-rate information (see also Lu et al., 2008).

The strategy of relying on an explanation's quality as a guide to its probability appears to be widespread, but is it warranted? This has been a topic of lively debate within philosophy, and for some properties of explanations, such as simplicity, within statistics and computer science as well (see Baker, 2010, for review). For example, some have suggested that simpler explanations avoid “overfitting” data and hence will generalize more effectively to novel cases (e.g., Forster, 2000). Others have suggested that an initial bias toward simplicity results in more efficient learning insofar as a learner is likely to modify hypotheses fewer times (e.g., Kelly, 2004). Breadth and unification have more direct ties to probability, as explanations that are broader or more general will, by definition, apply to more cases and could also have stronger evidential support (Myrvold, 2003). However, it is also possible that explanatory preferences result from cognitive limitations—for example, simpler explanations could be easier to process or remember—and that their influence on probability and inference is misguided. As one example, Pennington and Hastie (1992) find that presenting evidence to mock jurors in a chronological order, which makes the evidence easier to integrate into a coherent explanation for the events that occurred, systematically influences verdicts, in contrast to the predictions of most normative models.

In the cases considered so far, an explanation's quality is used as a guide to the probability of that *explanation*. Other findings suggest that the existence and quality of explanations can influence the judged probability of *what is being explained*. For example, Koehler (1991)(1991) reviews evidence that prompting participants to explain one outcome—say, why a Democrat might win the next election—boosts the probability assigned to that outcome. Similarly, explaining why a relationship holds—say, why people who are risk-averse might make better firefighters—increases the probability assigned to that relationship (e.g., Anderson & Sechler, 1986; see also Ross et al., 1977), as well as memory for the relevant correlation (Bower & Masling, unpublished data).

Thagard (1989) proposes a model of abductive inference that incorporates both of the influences just considered: how the quality of an explanation influences the degree of belief in that explanation *as well as* belief in what is explained. Thagard's approach, the Theory of Explanatory Coherence, involves seven principles that specify relationships between propositions and their consequences for belief. For example, the principles specify that propositions that participate in explanatory relationships (whether they do the explaining or are explained) are mutually reinforcing, as are those that provide analogous explanations. The model has been instantiated in a connectionist network and successfully used to model human judgments (e.g., Ranney & Thagard, 1988; Thagard,

2006), providing another source of evidence for the importance of explanatory evaluation in inference.

In addition to direct influences of explanatory evaluations on the probability assigned to an explanation and what it explains, explanations influence judgments in tasks that require assessing the probability of one claim in light of another. For example, Sloman (1994) provided participants with an initial claim, such as many secretaries "have a hard (p. 270) time financing a house" or "have bad backs," and asked them to evaluate the probability of a related claim, such as many furniture movers "have a hard time financing a house" or "have bad backs." When the claims supported a common explanation (e.g., moderate income for trouble financing a house), participants provided higher probability estimates than when the claims supported different explanations (e.g., a sedentary job versus heavy lifting for back problems; see also Rehder, 2006). This could be because explanations compete (as might be expected in a framework like Thagard's Explanatory Coherence), or because an explanation that applies to many cases is judged to be broader or more unifying and therefore of higher quality, resulting in a boost to the probability of what it explains (in keeping with the findings reviewed above).

The role of explanation in inference extends to a variety of additional judgments, including category membership (Rips et al., Chapter 11), decision making (LeBoeuf & Shafir, Chapter 16; see also Hastie & Pennington, 2000), and both legal (Spellman & Schauer, Chapter 36) and medical (Patel et al., Chapter 37) reasoning. For example, in the context of concept learning and categorization, explanatory knowledge has been shown to facilitate learning (Murphy & Allopenna, 1994; Williams & Lombrozo, 2010), influence judgments of the typicality of category members (Ahn, Marsh, Luhmann, & Lee, 2002; Murphy & Allopenna, 1994), and foster conceptual coherence (Murphy & Medin, 1985; Patalano, Chin-Parker, & Ross, 2006). In addition, the way in which a category's features are explained can influence the relative importance of those features in making judgments about category membership: Features that support more explanations (Sloman, Love, & Ahn, 1998), causally deeper explanations (Ahn, 1998; Ahn & Kim, 2000), or explanations that are privileged by the explanatory stance adopted—mechanistic or functional (T. Lombrozo, 2009)—will typically be weighted more heavily. Such findings illustrate the breadth and importance of explanation's effects.

Finally, although explanations likely benefit inference in many contexts, there are some well-documented cases in which explanatory judgments can lead people astray. For example, both children and adults tend to mistake explanations for *why* something is the case for evidence *that* something is the case (Glassner, Weinstock, & Neuman, 2005; Kuhn, 2001), especially when evidence is unavailable (Brem & Rips, 2000). People's ability to detect circularity in explanations is also imperfect (Baum, Danovitch, & Keil, 2008; Rips, 2002) and susceptible to extraneous factors, such as the presence of interesting but potentially irrelevant information (Weisberg, Keil, Goodstein, Rawson, & Gray, 2008). Finally, both children and adults tend to overestimate the accuracy and depth of their own explanations (Mills & Keil, 2004; Rozenblit & Keil, 2002). So while

there is little doubt that explanation has a considerable influence on a wide range of everyday inferences, this influence is quite heterogeneous in scope and consequences, and almost certainly in underlying mechanisms.

Distinguishing Explanation From Other Cognitive Processes

Explanation is closely related to a variety of cognitive processes, including but not limited to inductive reasoning, deductive reasoning, categorization, causal reasoning, analogical reasoning, and learning. Can and should explanation be distinguished from these extant areas of study? From one perspective, the answer is clearly “no.” Explanation could trigger or be triggered by these processes, and it is likely to share many common mechanisms. Isolating explanation could thus result in a mischaracterization of cognitive mechanisms and their coordination. But there are also compelling reasons to study explanation and abductive inference as phenomena in their own right. The study of explanation highlights important aspects of learning, reasoning, and representation that are obscured when explanation is undifferentiated from general causal reasoning or inference. This section identifies a few distinctions that help locate the study of explanation and its unique contributions.

First, are explanation and abductive inference simply kinds of causal reasoning? One compelling reason to avoid this identification is because it assumes that all explanations are causal (see earlier section on “The Structure of Explanations”). There are broad classes of explanations that most people acknowledge as noncausal, such as mathematical and logical explanations. There are also explanations involving causal systems that are arguably noncausal. For example, consider an explanation for why a 1-inch square peg won't fit into a 1-inch-diameter round hole. While causal factors are involved, it is arguably geometric facts that do the explanatory work (Putnam, 1975). The reverse also holds: Not all causal hypotheses are explanatory. It is the burden of a theory of explanation (p. 271) to account for why the big bang is a poor explanation for cultural differences in color preferences, and more generally why some and only some causes strike us as explanations for their effects. Identifying “explanatory” causes—those that figure in a selected explanation—could be akin to what we do when we identify “the” cause of an effect or distinguish causes from enabling conditions, but it is certainly not equivalent to causal inference or to the specification of all causal factors.

Second, is it worth distinguishing explanations and abductive inference from the broader and quite heterogeneous class of inductive or deductive inferences? Again, not all explanations are straightforward examples of inference, and not all inferences are explanatory. Consider a case in which the causes of an event are firmly established: A match was struck in the presence of oxygen, and a fire ensued. In such a case, the striking of the match is likely to be selected as the explanation for the fire, not the

presence of oxygen. This judgment requires an inference concerning which aspects of the causal structure are *explanatory*, but it need not be accompanied by an inference concerning which causes were present or type-level causal relationships, as these are already known. More important, explanatory inferences are a subset of everyday inferences. I can infer what the weather will be like based on a forecast, or what the 100th digit of pi is given the output of a computer program. But these inferences are not explanatory, and as a result the characteristics of abductive inference—such as a role for simplicity or breadth in determining the perceived probability of a hypothesis—need not apply.

Recognizing the unique aspects of explanation brings important questions into relief. For example, what is common between causal and noncausal explanations? Why do causal explanations privilege some kinds or aspects of causal structure over others? How do explanatory inferences go beyond inferences concerning causal and statistical structure? Do explanatory inferences involve unique mechanisms for evaluation? These questions can best be addressed through systematic investigation of the structure and functions of explanation, including new empirical findings to complement those reviewed here.

Conclusions

This chapter began by considering the role of explanation in cognition, and in particular why explanation is such a pervasive aspect of human experience and elicits such strong intuitions. The reviewed findings confirm that the ability to evaluate and be guided by explanations is already quite sophisticated in young children and persists throughout adulthood. Moreover, explanatory preferences are highly systematic and have important consequences for core cognitive capacities that span learning and inference.

While explanations are quite heterogeneous and almost certainly impact cognition through a variety of distinct mechanisms, effects of explanations typically share a few characteristics. In particular, the process of explaining appears to recruit relevant prior beliefs along with a suite of formal constraints in the form of explanatory preferences—that is, preferences for explanations that have particular characteristics, such as simplicity or the ability to subsume and unify disparate observations. These properties of explanations (the products) account for many of the consequences of explanation (the process), which range from quite beneficial to neutral or even detrimental.

Despite a focus on both the functional consequences of explanations and how explanations are evaluated, there have been few attempts to provide comprehensive computational- or algorithmic-level accounts of explanation. In the absence of computational-level accounts, explanations and explanatory inferences lack normative benchmarks against which to assess human performance. And in the absence of algorithmic-level accounts, the relationship between explanation and other cognitive

capacities remains largely opaque. Developing such accounts should thus be a high priority as research on explanation and abductive inference moves forward. The final section of the chapter identifies additional, promising avenues for future research.

Future Directions

Research on explanation is impressive in its scope and diversity, spanning cognitive psychology, social psychology, developmental psychology, education, philosophy, and beyond. However, distinct strands of research have proceeded almost independently, and until recently much of this work has not been unified under a common umbrella. A key direction for future research, then, involves the integration of different perspectives and findings across these areas. However, existing research raises more questions than it answers, and many aspects of explanations are only beginning to be explored. This section highlights a few questions that are ripe for further (p. 272) investigation. The first six have already been raised in the course of the review; the final four involve additional issues.

1. Are there different kinds of explanations? If so, in virtue of what are they all explanatory?
2. What are the functions of explanation? To what extent are systematic explanatory preferences attributable to these functions, and to what extent do they result as side effects of other cognitive processes?
3. Which explanations are sought, and why? In particular, how do the characteristics of the explainer's prior beliefs interact with the environment to determine what is deemed to require explanation?
4. What are the mechanisms by which explanation guides learning? How is the role of explanation influenced by feedback on the accuracy of explanations or from further observations or testimony? When is the role of explanation beneficial, and when is it detrimental?
5. What are the criteria employed in the evaluation of explanations, and what are the consequences for learning and inference? In particular, what are the conditions under which explanatory considerations—such as an explanation's simplicity—inform assessments of probability, and is this role for explanatory considerations ever warranted?
6. What are the prospects for a normative or computational-level theory of explanation and abductive inference? How will such a theory inform empirical research, and how should the theory be constrained by empirical findings?
7. How do social and pragmatic factors influence explanations and their consequences?

8. To what extent are there individual differences and cultural variation in explanatory preferences? What is the source of this variation, and what are the consequences for learning and inference?

9. To what extent can explanation be separated from language? While typical explanations are encoded in language, they need not be. For example, Baillargeon (2004) proposed that prelinguistic infants construct explanations, while Povinelli and Dunphy-Lelii (2001) devised a clever method for assessing "explanation" in chimpanzees. As a conceptual question, how can explanation be characterized without a commitment to some language-like manifestation? And as a methodological question, how can nonlinguistic explanations be studied empirically?

10. How are explanations generated? There has been much more research on how explanations are evaluated than on how they are generated. This is in part because explanation generation confronts many of the most difficult questions in cognitive science concerning the content and structure of general beliefs and how these are retrieved in particular contexts. How can the study of explanation generation be informed by research on representation and memory, and how might findings concerning the generation of explanations in turn contribute to these areas?

Acknowledgments

Sincere thanks to Joseph Austerweil, Thomas Griffiths, Ulrike Hahn, Keith Holyoak, G. Randolph Mayes, and Joseph Williams for comments on earlier drafts of this chapter, and NSF grant DRL-1056712 for support..

References

Ahn, W. (1998). Why are different features central for natural kinds and artifacts? *Cognition*, 69, 135–178.

Ahn, W., & Bailenson, J. (1996). Causal attribution as a search for underlying mechanisms: An explanation of the conjunction fallacy and the discounting principle. *Cognitive Psychology*, 31, 82–123.

Ahn, W., Brewer, W. F., & Mooney, R. J. (1992). Schema acquisition from a single example. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 391–412.

Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation vs. mechanism information in causal attribution. *Cognition*, 54, 299–352.

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

Ahn, W., Marsh, J. K., Luhmann, C. C., & Lee, K. (2002). Effect of theory-based feature correlations on typicality judgments. *Memory and Cognition*, 30, 107–118.

Ahn, W., Novick, L., & Kim, N. S. (2003). Understanding it makes it more normal. *Psychonomic Bulletin and Review*, 10, 746–752.

Amsterlaw, J., & Wellman, H. (2006). Theories of mind in transition: A microgenetic study of the development of false belief understanding. *Journal of Cognition and Development*, 7, 139–172.

Anderson, C. A., Krull, D. S., & Weiner, B. (1996). Explanations: Processes and consequences. In E.T. Higins & A.W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 271–296). New York: Guilford Press.

Anderson, C. A., & Sechler, E. S. (1986). Effects of explanation and counterexplanation on the development and use of social theories. *Journal of Personality and Social Psychology*, 50, 24–34.

Baker, A. (2010). Simplicity. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2010 ed.). URL Retrieved August 2011, from **http://plato.stanford.edu/archives/spr2010/entries/simplicity/** (p. 273)

Baillargeon, R. (2004). Infants' physical world. *Current Directions in Psychological Science*, 13, 89–94.

Baram-Tsabari, A., & Yarden, A. (2005). Characterizing children's spontaneous interests in science and technology. *International Journal of Science Education*, 27, 803–826.

Baum, L. A., Danovitch, J. H., & Keil, F. C. (2008). Children's sensitivity to circular explanations. *Journal of Experimental Child Psychology*, 100, 146–155.

Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. London: Routledge.

Berlyne, D. E. (1954). A theory of human curiosity. *British Journal of Psychology*, 45, 180–191.

Bonawitz, E.B. & Lombrozo, T. (in press). Occam's rattle: children's use of simplicity and probability to constrain inference. *Developmental Psychology*.

Brem, S. K., & Rips, L. J. (2000) Explanation and evidence in informal argument. *Cognitive Science*, 24, 573–604.

Brown, A. L., & Kane, M. J. (1988). Preschool children can learn to transfer: Learning to learn and learning from example. *Cognitive Psychology*, 20, 493–523.

Burch, R. (2010). Charles Sanders Peirce. In E. N. Zalta (Ed.), The Stanford encyclopedia of philosophy (Fall 2010 ed.). Retrieved August 2011, from **http://plato.stanford.edu/archives/fall2010/entries/peirce/**

Callanan, M. A., & Oakes, L. (1992). Preschoolers' questions and parents' explanations: Causal thinking in everyday activity. *Cognitive Development*, 7, 213–233.

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

Casler, K., & Kelemen, D. (2008). Developmental continuity in the teleo-functional explanation: Reasoning about nature among Romanian Romani adults. *Journal of Cognition and Development*, 9, 340–362.

Chi, M. T. H. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology* (pp. 161–238). Hillsdale, NJ: Erlbaum.

Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, 1, 73–105.

Chi, M. T. H., Bassok, M., Lewis, M., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13, 145–182.

Chi, M. T. H., de Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18, 439–477.

Chin-Parker, S., & Bradner, A. (2010). Background shifts affect explanatory style: How a pragmatic theory of explanation accounts for background effects in the generation of explanations. *Cognitive Processing*, 11, 227–249.

Chouinard, M. (2007). Children's questions: A mechanism for cognitive development. *Monographs of the Society for Research in Child Development*, 72, 1–57.

Craik, K. (1943). *The nature of explanations*. Cambridge, England: Cambridge University Press.

Craver, C. (2007). *Explaining the brain: What a science of the mind-brain could be*. New York: Oxford University Press.

Crowley, K., & Siegler, R. S. (1999). Explanation and generalization in young children's strategy learning. *Child Development*, 70, 304–316.

Darden, L. (2006). *Reasoning in biological discoveries: Essays on mechanisms, interfiled relations, and anomaly resolution*. Cambridge, England: Cambridge University Press.

Dawes, R.M. (1999). A message from psychologists to economists: Mere predictability doesn't matter like it should (without a god story appended to it). *Journal of Economic Behavior & Organization*, 39, 29–40.

DeJong, G. (2004). Explanation-based learning. In A. Tucker (Ed.), *Computer science handbook* (2nd ed., pp. 68–1–68–20).

Dennett, D. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin*, 99, 3–19.

Fonseca, B. A., & Chi, M. T. H. (2010). Instruction based on self-explanation. In R. Mayer & P. Alexander (Eds.), *The handbook of research on learning and instruction* (pp. 296–321). New York: Routledge Press.

Forster, M. R. (2000). Key concepts in model selection – performance and generalizability. *Journal of Mathematical Psychology*, 44, 205–231.

Frazier, B. N., Gelman, S. A., & Wellman, H. M. (2009). Preschoolers' search for explanatory information within adult-child conversation. *Child Development*, 80, 1592–1611.

Friedman, M. (1974). Explanation and scientific understanding. *Journal of Philosophy*, 71, 5–19.

Garfinkel, A. (1990). *Forms of explanation: Rethinking the questions in social theory*. New Haven, CT: Yale University Press.

Glassner, A., Weinstock, M., & Neuman, Y. (2005). Pupils' evaluation and generation of evidence and explanation in argumentation. *British Journal of Educational Psychology*, 75, 105–118

Goodman, N. D., Baker, C. L., Bonawitz, E. B., Mansinghka, V. K., Gopnik, A., Wellman, H., . . . Tenenbaum, J. B. (2006). Intuitive theories of mind: A rational approach to false belief. In R. Sun, N. Miyake, C. Schunn, & S. Lane (Eds.), *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society* (pp. 1382–1387).

Gopnik, A. (2000). Explanation as orgasm and the drive for causal knowledge: The function, evolution, and phenomenology of the theory-formation system. In F. Keil & R. A. Wilson (Eds.), *Explanation and cognition* (pp. 299–324). Cambridge, MA: MIT Press.

Gopnik, A., & Meltzoff, A. (1997). *Words, thoughts and theories*. Cambridge, MA: MIT Press.

Graesser, A. C., Langston, M. C., & Baggett, W. B. (1993). Exploring information about concepts by asking questions. In G. V. Nakamura, R. M. Taraban, & D. Medin (Eds.), *The*

psychology of learning and motivation. Vol. 29: Categorization by humans and machines (pp. 411–436). Orlando, FL: Academic Press.

Greif, M., Kemler-Nelson, D., Keil, F. C., & Guiterrez, F. (2006). What do children want to know about animals and artifacts? Domain-specific requests for information. *Psychological Science*, 17, 455–459.

Harman, G. (1965). The inference to the best explanation. *Philosophical Review*, 74, 88–95.

Hart, H. L. A., & Honoré, T. (1985). Causation in the law (2nd ed.). New York: Oxford University Press.

Hastie, R., & Pennington, N. (2000) Explanation-based decision making. In T. Connolly, H. R. Arkes, & K. R. Hammond (Eds.), *Judgment and decision making: An interdisciplinary reader* (2nd ed., pp. 212–228). Cambridge, England: Cambridge University Press. (p. 274)

Heider, F. (1958). *The psychology of interpersonal relations*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hempel, C. (1965). *Aspects of scientific explanation and other essays in the philosophy of science*. New York: Free Press.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, 15, 135–175.

Heussen, D. S. (2010). When functions and causes compete. *Thinking and Reasoning*, 16, 233–250.

Hickling, A. K., & Wellman, H. M. (2001). The emergence of children's causal explanations and theories: Evidence from everyday conversation. *Developmental Psychology*, 37, 668–683.

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, 107, 65–81.

Hilton, D. J. (1996). Mental models and causal explanation: Judgments of probable cause and explanatory relevance. *Thinking and Reasoning*, 2, 273–308.

Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, 93, 75–88.

Hood, L., & Bloom, L. (1979). What, when, and how about why: A longitudinal study of the early expressions of causality. *Monographs of the Society for Research in Child Development*, 44(6, serial no. 181).

Hooper, S. (1992). Effects of peer interaction during computer-based mathematics instruction. *Journal of Educational Research*, 85, 180–189.

Isaacs, N. (1930). Children's “why” questions. In S. Isaacs (Ed.), *Intellectual growth in young children* (pp. 291–349). London: Routledge.

Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review*, 111, 640–661.

Keil, F. C. (1992). The origins of an autonomous biology. In M. R. Gunnar & M. Maratsos (Eds.), *Modularity and constraints in language and cognition. Vol. 25: Minnesota Symposium on Child Psychology* (pp. 103–138). Hillsdale, NJ: Erlbaum.

Keil, F. C. (1994). The birth and nurturance concepts by domains: The origins of concepts of living things. In L. A. Hirschfeld & S. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 234–254). Cambridge, England: Cambridge University Press.

Keil, F. C. (1995). The growth of causal understanding of natural kinds. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multi-disciplinary debate* (pp. 234–262). Oxford, England: Clarendon Press.

Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Science*, 7, 368–373.

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, 57, 227–254.

Keil, F. C., & Wilson, R. A. (2000). *Explanation and cognition*. Cambridge, MA: The MIT Press.

Kelemen, D. (1999) Function, goals and intention: Children's teleological reasoning about objects. *Trends in Cognitive Science*, 3, 461–468.

Kelemen, D., & Rosset, E. (2009). The human function compunction: Teleological explanation in adults. *Cognition*, 111, 138–143.

Kelly, K. (2004). Justification as truth-finding efficiency: How Ockham's Razor works. *Minds and Machines*, 14, 485–505.

Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher & W. Salmon (Eds.), *Scientific explanation* (pp. 410–505). Minneapolis: University of Minnesota Press.

Koehler, D. J. (1991). Explanation, imagination, and confidence in judgment. *Psychological Bulletin*, 110, 499–519.

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning*. Cambridge, MA: MIT Press.

Krull, D. S., & Anderson, C. A. (2001). Explanation, cognitive psychology of. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences* (Vol. 8, pp. 5150–5154). Oxford, England: Elsevier.

Kuhn, D. (2001). How do people know? *Psychological Science*, 12, 1–8.

Kuhn, D., & Katz, J. (2009). Are self-explanations always beneficial? *Journal of Experimental Child Psychology*, 103, 386–394.

Lagnado, D. (1994). *The psychology of explanation: A Bayesian approach*. Unpublished Masters thesis. Schools of Psychology and Computer Science, University of Birmingham, England.

Legare, C. H. (in press). Exploring explanation: Explaining inconsistent evidence informs exploratory, hypothesis-testing behavior in young children. *Child Development*.

Legare, C. H., Gelman, S. A., & Wellman, H. M. (2010). Inconsistency with prior knowledge triggers children's causal explanatory reasoning. *Child Development*, 81, 929–944.

Legare, C. H., Wellman, H. M., & Gelman, S. A. (2009). Evidence for an explanation advantage in naïve biological reasoning. *Cognitive Psychology*, 58, 177–194.

Lipton, P. (2004). *Inference to the best explanation*. New York: Routledge.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, 10, 464–470.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, 55, 232–257.

Lombrozo, T. (2009). Explanation and categorization: How “why?” informs “what?” *Cognition*, 110, 248–253.

Lombrozo, T. (2010a). From conceptual representations to explanatory relations. *Behavioral and Brain Sciences*, 33, 218–219.

Lombrozo, T. (2010b). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, 61, 303–332.

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, 99, 167–204.

Lombrozo, T., Kelemen, D., & Zaitchik, D. (2007). Inferring design: Evidence of a preference for teleological explanations in patients with Alzheimer's Disease. *Psychological Science*, 18, 999–1006.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, 115, 955–982.

Machery, E. (2009). *Doing without concepts*. New York: Oxford University Press.

Mackie, J. L. (1980). *The cement of the universe: A study of causation*. New York: Oxford University Press.

Magnani, L. (2001), *Abduction, reason, and science. Processes of discovery and explanation*. New York: Kluwer Academic/Plenum Publishers.

Malle, B. F. (2004). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Cambridge, MA: MIT Press.

Margolis, E. (1995). The significance of the theory analogy in the psychological study of concepts. *Mind and Language*, 10, 45–71.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. New York: W. H. Freeman. (p. 275)

Mills, C., & Keil, F. C. (2004). Knowing the limits of one's understanding: The development of an awareness of an illusion of explanatory depth. *Journal of Experimental Child Psychology*, 87, 1–32.

Miyake, N., & Norman, D. A. (1979). To ask a question one must know enough to know what is not known. *Journal of Verbal Learning and Verbal Behavior*, 18, 357–364.

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.

Myrvold, W. C. (2003). A Bayesian account of the virtue of unification. *Philosophy of Science*, 70, 399–423.

Needham, D. R., & Begg, I. M. (1991). Problem-oriented training promotes spontaneous analogical transfer: Memory-oriented training promotes memory for training. *Memory and Cognition*, 19, 543–557.

Patalano, A. L., Chin-Parker, S., & Ross, B. H. (2006). The importance of being coherent: Category coherence, cross-classification, and reasoning. *Journal of Memory and Language*, 54, 407–424.

Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the story-model for juror decision making. *Journal of Personality and Social Psychology*, 62, 189–206.

Pennington, N., & Hastie, R. (1993). The story model for juror decision making. In R. Hastie (Ed.), *Inside the juror: The psychology of juror decision making* (pp. 192–221). Cambridge, England and New York: Cambridge University Press.

Peterson, C., & Slaughter, V. (2003). Opening windows into the mind: Mothers' preferences for mental state explanations and children's theory of mind. *Cognitive Development*, 18, 399–429.

Piaget, J. (1929). *The child's conception of the world*. London: Routledge & Kegan Paul.

Povinelli, D. J., & Dunphy-Lelii, S. (2001). Do chimpanzees seek explanations? Preliminary comparative investigation. *Canadian Journal of Experimental Psychology*, 55, 185–193.

Prasada, S., & Dillingham, E. M. (2006). Principled and statistical connections in common sense conception. *Cognition*, 99, 73–112.

Prasada, S., & Dillingham, E. M. (2009). Representation of principled connections: A window onto the formal aspect of common sense conception. *Cognitive Science*, 33, 401–448.

Preston, J., & Epley, N. (2005). Explanations versus applications: The explanatory power of valuable beliefs. *Psychological Science*, 16, 826–832.

Putnam, H. (Ed.) (1975). Philosophy and our mental life. In *Mind, language and reality: Philosophical papers* (Vol. 2, pp. 291–303). Cambridge, England: Cambridge University Press.

Railton, P. (1981). Probability, explanation, and information. *Synthese*, 48, 233–256.

Ranney, M., & Thagard, P. (1988). Explanatory coherence and belief revision in naive physics. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society* (pp. 426–432). Hillsdale, NJ: Erlbaum.

Read, S. J., & Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, 65, 429–447.

Rehder, B. (2006). When causality and similarity compete in category-based property induction. *Memory and Cognition*, 34, 3–16.

Rips, L. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21–59). Cambridge, England: Cambridge University Press.

Rips, L. J. (2002). Circular reasoning. *Cognitive Science*, *26*, 767–795.

Rittle-Johnson, B. (2006). Promoting transfer: The effects of direct instruction and self-explanation. *Child Development*, *77*, 1–15.

Roscoe, R. D., & Chi, M. T. H. (2008). Tutor learning: The role of explaining and responding to questions. *Instructional Science*, *36*(4), 321–350.

Ross, J., & Cousins, J. B. (1995) Giving and receiving explanations in cooperative learning groups. *Alberta Journal of Educational Research*, *41*, 104–122.

Ross, L., Lepper, M. R., Strack, F., & Steinmetz, J. (1977). Social explanation and social expectation: Effects of real and hypothetical explanations on subjective likelihood. *Journal of Personality and Social Psychology*, *35*, 817–829.

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

Salmon, W. (1984). *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.

Salmon, W. (1989). *Four decades of scientific explanation.* Minneapolis: University of Minnesota Press.

Shtulman, A. (2006). Qualitative differences between naive and scientific theories of evolution. *Cognitive Psychology*, *52*, 170–194.

Shultz, T. R. (1980). Rules of causal attribution. *Monographs of the Society for Research in Child Development*, *47*, 1–51.

Siegler, R. S. (1995). How does change occur: A microgenetic study of number conservation. *Cognitive Psychology*, *28*, 225–273.

Siegler, R. S. (2002). Microgenetic studies of self-explanations. In N. Granott & J. Parziale (Eds.), *Microdevelopment: Transition processes in development and learning* (pp. 31–58). New York: Cambridge University Press.

Sloman, S. A. (1994). When explanations compete: The role of explanatory coherence on judgments of likelihood. *Cognition*, *52*, 1–21.

Sloman, S. A., Love, B. C., & Ahn, W. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, *22*, 189–228.

Strevens, M. (2008). *Depth: An account of scientific explanation.* Cambridge, MA: Harvard University Press.

Sully, J. (1900). *Studies of childhood*. New York: D. Appleton & Company.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, *12*, 435–467.

Thagard, P. (2000). Probabilistic networks and explanatory coherence. *Cognitive Science Quarterly*, *1*, 93–116.

Thagard, P. (2006). Evaluating explanations in science, law, and everyday life. *Current Directions in Psychological Science*, *15*, 141–145.

Trout, J. D. (2002). Scientific explanation and the sense of understanding. *Philosophy of Science*, *69*, 212–233.

Trout, J. D. (2008). Seduction without cause: Uncovering explanatory neurophilia. *Trends in Cognitive Sciences*, *12*, 281–282.

van Fraassen, B. C. (1980). *The scientific image.* Oxford, England: Oxford University Press.

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience*, *20*(3), 470–477.

Wellman, H. M. (2011). Reinvigorating explanations for the study of early cognitive development. *Child Development Perspectives*, *5*(1), 33–38. (p. 276)

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, *43*, 337–375.

Wellman, H. M., & Liu, D. (2007). Causal reasoning as informed by the early development of explanations. In A. Gopnik & L. E. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 261–279). New York: Oxford University Press.

Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science*, *34*, 776–806.

Williams, J. J., Lombrozo, T., & Rehder, B. (2010). Why does explaining help learning? Insight from an explanation impairment effect. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2906–2911). Austin, TX: Cognitive Science Society.

Wittwer, J., Nuckles, M., & Renkl, A. (2008). Is underestimation less detrimental than overestimations? The impact of experts' beliefs about a layperson's knowledge on learning and question asking. *Instructional Science*, *36*, 27–52.

Wittwer, J., & Renkl, A. (2008). Why instructional explanations often do not work: A framework for understanding the effectiveness of instructional explanations. *Educational Psychologist*, *43*, 49–64.

Woodward, J. (2003). *Making things happen: A theory of causal explanation.* Oxford, England: Oxford University Press.

Woodward, J. (2010). Scientific explanation. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2010 ed.). Retrieved August 2011, from **http://plato.stanford.edu/archives/spr2010/entries/scientific-explanation/**

Wright, L. (1976). *Teleological explanations.* Berkeley: University of California Press.

Tania Lombrozo

Department of Psychology University of California, Berkeley Berkeley, California, USA