Lombrozo, T. & Vasilyeva. N. (Forthcoming). Causal explanation. In M. Waldmann (Ed.), *Oxford Handbook of Causal Reasoning*. Oxford, UK: Oxford University Press.

Causal Explanation

Tania Lombrozo & Nadya Vasilyeva

Department of Psychology, University of California, Berkeley

Abstract

Explanation and causation are intimately related. Explanations often appeal to causes, and causal claims are often answers to implicit or explicit questions about why or how something occurred. In this chapter we consider what research on explanation can tell us about causal reasoning. In particular, we review an emerging body of work suggesting that explanatory considerations – such as the simplicity or scope of a causal hypothesis – can systematically influence causal inference and learning. We also discuss proposed distinctions among types of explanations and review their differential effects on causal reasoning and representation. Finally, we consider the relationship between explanations and causal mechanisms and raise important questions for future research.

Introduction

A doctor encounters a patient: why does she have a fever and a rash? An engineer investigates a failure: why did the bridge collapse? A parent wonders about her child: why did she throw a tantrum? In each of these cases, we seek an explanation for some event – an explanation that's likely to appeal to one or more antecedent *causes*. The doctor might conclude that a virus *caused* the symptoms, the engineer that defects in cast iron *caused* the bridge collapse, and the parent that the toy's disappearance *caused* the tantrum.

Not all explanations are causal, and not all causes are explanatory. Explanations in mathematics, for example, are typically taken to be non-causal, and many causal factors are either not explanatory at all, or only explanatory under particular circumstances. (Consider, for instance, appealing to the big bang as an explanation for today's inflation rates, or the presence of oxygen as an explanation for California wildfires.) Nonetheless, causation and explanation are closely related, with many instances of causal reasoning featuring explanations and explanatory considerations, and many instances of abductive inference and explanation appealing to causes and causal considerations. The goal of the present chapter is to identify some of the connections between explanation and causation, with a special focus on how the study of explanation can inform our understanding of causal reasoning.

The chapter is divided into five sections. In the first three, we review an emerging body of work on the role of explanation in three types of causal reasoning: drawing inferences about the causes of events, learning novel causal structures, and assigning causal responsibility. In the fourth section, we consider different kinds of explanations, including a discussion of whether each kind is properly "causal" and how different kinds of explanations can differentially influence causal judgments. In the fifth section, we focus on causal explanations that appeal to mechanisms, and consider the relationship between explanation, causal claims, and mechanisms. Finally, we conclude with some important questions for future research.

Causal Inference & Inference to the Best Explanation

Consider a doctor who infers, on the basis of a patient's symptoms, that the patient has a particular disease – one known to cause that cluster of symptoms. We'll refer to such instances of causal reasoning as "causal inference," and differentiate them from two other kinds of causal reasoning that we'll discuss in subsequent sections: causal learning (which involves learning about novel causes and relationships at the type level) and assigning causal responsibility (which involves attributing an effect to one or more causes, all of which could have occurred and contributed to the effect).

How might explanation influence causal inference? One possibility is that people engage in a process called "inference to the best explanation" (IBE). IBE was introduced into the philosophical literature by Gilbert Harman in a 1965 paper, but the idea is likely older, and closely related to what is sometimes called "abductive inference" (Douven, 2011; Lombrozo, 2012; Peirce, 1955). The basic idea is that one infers that a hypothesis is likely to be true based on the fact that it *best explains* the data. To borrow vocabulary from another influential philosopher of explanation, Peter Lipton, one uses an explanation's "loveliness" as a guide to its "likeliness" (Lipton, 2004).

A great deal of work has aimed to characterize how people go about inferring causes from patterns of evidence (Cheng, 1997; Cheng & Novick, 1990, 1992; Griffiths &

Tenenbaum, 2005; Kelley, 1973; Perales & Shanks, 2003; Shanks & Dickinson, 1988; Waldmann & Hagmayer, 2001; see Buehner, 2005; Holyoak & Cheng, 2011; Waldmann & Hagmayer, 2013, for reviews), and this work is summarized in other chapters of this volume (see the Theories of Causal Cognition section, and the chapter by Meder & Mayrhofer on diagnostic reasoning, this volume). Thus a question that immediately presents itself is whether IBE is distinct from the kinds of inference these models typically involve, such as analyses of covariation or Bayesian inference. For most advocates of IBE, the answer is "yes": IBE is a distinct inferential process, where the key commitment is that explanatory considerations play a role in guiding judgments. These considerations can include the simplicity, scope, or other "virtues" of the explanatory hypotheses under consideration.

To provide evidence for IBE as a distinctly explanatory form of inference, it's thus important to identify explanatory virtues, and to demonstrate their role in inference. The most direct evidence of this form comes from research on simplicity (Bonawitz & Lombrozo, 2012; Lombrozo, 2007; Pacer & Lombrozo, in prep), scope (Khemlani, Sussman, & Oppenheimer, 2011), and explanatory power (Douven & Schupbach, 2015a, 2015b). We focus on this research for the remainder of the section.

In one study from Lombrozo (2007), participants learned novel causal structures describing the relationships between diseases and symptoms on an alien planet. For example, the conjunction of two particular symptoms – say sore minttels and purple spots – could be explained by appeal to a single disease that caused both symptoms, Tritchet's syndrome, or by appeal to the conjunction of two diseases that each caused one symptom, Morad's disease and a Humel infection. Lombrozo set out to test whether participants would favor the explanation that was simpler in the sense that it invoked a single common cause over two independent causes, and whether they would do so even when probabilistic evidence, in the form of disease baserates, favored the more complex explanation. Lombrozo found that participants' explanation choices were a function of both simplicity and probability, with a substantial proportion of participants selecting the simpler explanation even when it was less likely than the complex alternative. This is consistent with the idea that an explanation's "loveliness" – in this case, its simplicity – is used as a basis for inferring its "likeliness."

In subsequent work, Bonawitz and Lombrozo (2012) replicated the same basic pattern of results with 5-year-old children in a structurally parallel task: children observed a toy generate two effects (a light and a spinning fan), and had to infer whether one block (which generated both effects) or two blocks (which each generated one effect) fell into the toy's activator bin. In this case, probabilistic information was manipulated across participants by varying the number of blocks of each type and the process by which they fell into the bin. Interestingly, adults did not show a preference for simplicity above and beyond probability in this task, while the 5-year-olds did. Bonawitz and Lombrozo suggest that in the face of probabilistic uncertainty – of the kind that's generated by a more complex task like the alien diagnosis problems used in Lombrozo (2007) – adults rely on explanatory considerations such as simplicity to guide assessments of probability. But when a task involves a transparent and seemingly deterministic causal system, and when the numbers involved are small (as was the case for the task developed for young children in Bonawitz and Lombrozo, 2012), adults may engage in more explicit probabilistic reasoning, and bypass explanatory considerations altogether. Consistent with this idea, adults in Lombrozo (2007) also ceased to favor simplicity when they were explicitly told that the complex hypothesis was most likely to be true.

In more recent work, Pacer and Lombrozo (2015) provide a more precise characterization of how people assess an explanation's simplicity. They differentiate two intuitive metrics for causal explanations, both of which are consistent with prior results: "count simplicity," which involves counting the number of causes invoked in an explanation, and "root simplicity," which involves counting the number of *unexplained* causes invoked in an explanation. For example, suppose that Dr. Count explains a patient's symptoms by appeal to pneumonia and sarcoma – two diseases. And that Dr. Root explains the symptoms by appeal to pneumonia, sarcoma, *and HIV*, where HIV is a cause (or at least a contributing factor) for both pneumonia and sarcoma. Dr. Root has invoked more causes than Dr. Count (three versus two), and so her explanation is less simple according to count simplicity. But Dr. Root has explained the symptoms by appeal to only one *unexplained* cause (HIV) as opposed to Dr. Count's two (pneumonia and sarcoma), so her explanation is simpler according to root simplicity. Extending the basic method developed by Lombrozo (2007), Pacer and Lombrozo found strong evidence that people favor explanations with low root simplicity (above and beyond what's warranted on the basis of the frequency information with which they were provided), but no evidence that people are sensitive to count simplicity. By using appropriate causal structure, they were able to rule out alternative explanations for these results (e.g., that people prefer explanations that involve intervening variables).

These findings suggest that in drawing causal inferences, people do not simply engage in probabilistic inference on the basis of frequency information. In addition to frequency information, they use explanatory considerations (in this case, low root simplicity) to guide their judgments, at least in the face of probabilistic uncertainty. The findings therefore suggest that IBE plays a role in inferences concerning causal events. But is this effect restricted to simplicity, or do other explanatory considerations play a role as well? Research to date supports a role for two additional factors: *narrow latent scope* and *explanatory power*.

An explanation's "latent scope" refers to the number of unverified effects that the explanation predicts. For example, an observed symptom could be explained by appeal to a disease that predicts that single symptom, or by appeal to a disease that additionally predicts an effect that has not yet been tested for and is hence unobserved (e.g., whether the person has low blood levels of some mineral). In this case, the former explanation has narrower latent scope. Khemlani, Sussman, and Oppenheimer (2011) found that people favor explanations with narrow latent scope, even if the two diseases are equally prevalent. Importantly, they also find that latent scope affects probability estimates: explanations with narrow latent scope are judged more likely than those with broader latent scope (see also Johnson, Johnston, Toig, & Keil, 2014, for evidence that explanatory scope informs causal strength inferences, and Johnston, Johnson, Koven, & Keil, 2015, for evidence of latent scope bias in children). Thus latent scope appears to be among the cues to explanatory "loveliness" that affect the perceived "likeliness" of explanatory hypotheses.

Finally, recent work by Douven and Schupbach (2015a, 2015b) provides further evidence of a role for explanatory considerations in inference, with hints that the relevant consideration is "explanatory power." Employing a quite different paradigm, Douven and Schupbach demonstrate that people's explanatory judgments better predict their estimates of posterior probability than do objective probabilities on their own. In a study reported in Douven and Schupbach (2015a), participants observed ten balls successively drawn from one of two urns, which was selected by a coin flip. One urn contained 30 black balls and 10 white balls, and the other contained 15 black balls and 25 white ones. After each draw, participants were asked to consider the evidence so far, and to rate the "explanatory goodness" for each of two hypotheses: the hypothesis that the balls were drawn from the 30/10 urn, or the hypothesis that the balls were drawn from the 15/25 urn. Participants were also asked to estimate a posterior probability for each hypothesis after each draw. In a series of models, Douven and Schupbach tested whether people's judgments of the explanatory "goodness" of each hypothesis improved model predictions of their subjective posterior probabilities, above and beyond the objective posteriors calculated on the basis of the data presented to each participant. They found that models incorporating these explanatory judgments outperformed alternatives, even when appropriately penalized for using additional predictors.

Douven and Schupbach's (2015a) results suggest that explanatory considerations do inform assessments of probability, and that these considerations diverge from posterior probability. However, the findings don't pinpoint the nature of the explanatory considerations themselves. On what basis were participants judging one hypothesis more or less explanatory than the other? Additional analyses of these data, reported in Douven and Schupbach (2015b), provide some hints: models that took into account some measure of "explanatory power" – computed on the basis of the objective probabilities – outperformed the basic model that only considered posteriors. The best-performing model employed a measure based on Good (1960) that roughly tracks *confirmation*: it takes the log of the ratio of the probability of the data given the hypothesis to the probability of the data. In other work, Schupbach (2011) finds evidence that people's judgments of explanation "goodness" are related to another measure of explanatory power, proposed by Schupbach and Sprenger (2011), which is also related to Bayesian measures of confirmation.

These findings suggest that explanatory considerations – in the form of root simplicity, latent scope, and explanatory power – inform causal inference, and in so doing reveal something potentially surprising: that while people's responses to evidence are systematic, they don't (always) lead to causal inferences that track the posterior probabilities of each causal hypothesis. This not only supports a role for explanatory considerations in causal inference, but also challenges the idea that identifying causes to *explain* effects is essentially a matter of conditionalizing on the effects to infer the most likely cause. Further challenging this idea, Pacer, Williams, Chen, Lombrozo, and Griffiths (2013) compare judgments of explanatory goodness from human participants to those generated by four distinct computational models of explanation in causal Bayesian networks, and find that models that compute measures of evidence or information considerably outperform those that compute more direct measures of (posterior) probability.

In sum, there's good evidence that people engage in a process like IBE when drawing inferences about causal events: they use explanatory considerations to guide their assessments of which causes account for observed effects, and of how likely candidate hypotheses are to be true. The most direct evidence to date concerns root simplicity, latent scope, and explanatory power, but there's indirect evidence that other explanatory considerations, such as coherence, completeness, and manifest scope, may play a similar role (Pennington & Hastie, 1988; Read & Marcus-Newhall, 1993; Preston & Epley, 2005; Thagard, 1989; Williams & Lombrozo, 2010).

Before concluding this section on IBE in causal inference, it's worth considering the normative implications of this work. It's typically assumed that Bayesian updating provides the normatively correct procedure for revising belief in causal hypotheses in light of the evidence. Do the findings reported in this section describe a true departure from Bayesian inference, and therefore a systematic source of *error* in human judgment? This is certainly one possibility. For example, it could be that IBE describes an imperfect algorithm by which people *approximate* Bayesian inference. If this is the case, it becomes an interesting project to spell out when and why explanatory considerations ever *succeed* in approximating more direct probabilistic inference.

There are other possibilities, however. In particular, an appropriately specified Bayesian model could potentially account for these results. In fact, some have argued that IBE-like inference could simply fall out of hierarchical Bayesian inference with suitably assigned priors and likelihoods (Henderson, 2013), in which case there could be a justified, Bayesian account of this behavior. It could also be that the Bayesian models implicit in the comparisons between people's judgments and posterior probabilities fail to describe the inference that people are actually making. In their chapter in this volume on diagnostic reasoning for example, Meder and Mayrhofer (this volume) make the important point that there can be more than one "Bayesian" model for a given inference, and in fact find different patterns of inference for models that make different assumptions when it comes to elemental diagnostic reasoning: inferring the value of a single binary cause from a single binary effect, which has clear parallels to the cases considered here. In particular, they argue for a model that takes into account uncertainty in causal structures over one that simply computes the empirical conditional probability of a cause given an effect. Similarly, it could be that the "departures" from Bayesian updating observed here reflect the consequences of a Bayesian inference that involves more than a straight calculation of posteriors.

Finally, some argue that IBE corresponds to a distinct but normatively justifiable *alternative* to Bayesianism (e.g., Douven & Schupbach, 2015). In particular, while Bayesian inference may be the best approach for minimizing expected inaccuracy in the long run, it could be that a process like IBE dominates Bayesian inference when the goal is, say, to get things mostly right in the short term, or to achieve some other aim (Douven, 2013). It could also be that explanations trade off other considerations against accuracy, such as the ease with which the explanation can be communicated, remembered, or used in subsequent processing. These are all important possibilities to explore in future research.

Causal Learning & The Process of Explaining

Consider a doctor who, when confronted with a recurring pattern of symptoms, posits a previously undocumented disease, or a previously unknown link between some pathogen and those symptoms. In each case, the inference involves a change in the doctor's beliefs about the causal structure of the world, not only about the particular patient's illness. This kind of inference, which we'll refer to as *causal model learning*, differs from the kinds of causal inferences considered in the preceding section in that the learner posits a novel cause or causal relation, not (only) a new token of a known type.

Just as explanatory considerations can influence causal inference, it's likely that a process like IBE can guide causal model learning. In fact, "Occam's Razor," the classic admonition against positing unnecessary types of entities (Baker, 2013), is typically formulated and invoked in the context of positing novel types, not tokens of known types. However, research to date has not (to our knowledge) directly explored IBE in the context of causal model learning. Doing so would require assessing whether novel causes or causal relations are more likely to be inferred when they provide better explanations.

What we do know is that engaging in explanation – *the process* – can affect the course of causal learning. In particular, a handful of studies with preschool-aged children suggest that being prompted to explain, even without feedback on the content or quality of explanations, can promote understanding of number conservation (Siegler, 1995) and of physical phenomena (e.g., a balance beam, Pine & Siegler, 2003), and recruit causal beliefs that aren't invoked spontaneously to guide predictions (Amsterlaw & Wellman, 2006; Bartsch & Wellman, 1989; Legare, Wellman, & Gelman, 2009). Prompts to explain can also accelerate children's understanding of false belief (Wellman & Lagattuta, 2004; Amsterlaw & Wellman, 2006; see Wellman & Liu, 2007 and Wellman, 2011 for reviews), which requires a revision from one causal model of behavior to a more complex model involving an unobserved variable (belief) and a causal link between beliefs and behavior (e.g., Goodman et al., 2006). Finally, there's evidence that prompting children to explain can lead them to preferentially learn about and remember causal mechanisms over causally-irrelevant perceptual details (Legare & Lombrozo, 2014), and that prompting children to explain makes them more likely to generalize internal parts and category membership from some objects to others on the basis of shared causal affordances as opposed to perceptual similarity (Walker, Lombrozo, Legare, & Gopnik, 2014; see also Muentener & Schulz, this volume, for more on children's causal learning).

To better understand the effects of explanation on children's causal learning, Walker, Lombrozo, Williams, Rafferty, and Gopnik (2015) set out to isolate effects of explanation on two key factors in causal learning: evidence and prior beliefs. Walker et al. used the classic "blicket detector" paradigm (Gopnik & Sobel, 2000), in which children observe blocks placed on a machine, where some of the blocks make the machine play music. Children have to learn which blocks activate the machine, which can involve positing a novel kind corresponding to a subset of blocks, and/or positing a novel causal relationship between those blocks (or some of their features) and the machine's activation.

In Walker et al.'s studies, 5-year-old children observed eight blocks successively placed on the machine, where four activated the machine and four did not. Crucially, half the children were prompted to *explain* after each observation ("Why did [didn't] this block make my machine play music?"), and the remaining children, in the control condition, were asked to report the outcome ("What happened to my machine when I put this block on it? Did it play music?"). This control task was intended to match the explanation condition in eliciting a verbal response and drawing attention to the relaionship between each block and the machine, but without requiring that the child explain.

Across studies, Walker et al. varied the properties of the blocks to investigate whether prompting children to explain made them more likely to favor causal hypotheses that were more consistent with the data (i.e., one hypothesis accounted for 100% of observations and the other for 75%) and/or more consistent with prior beliefs (i.e., one hypothesis involved heavier blocks activating the machine, which matched children's initial asumptions; the other involved blocks of a given color activating the machine). When competing causal hypotheses were matched in terms of prior beliefs but varied in the evidence they accounted for, children who were prompted to explain were significantly more likely than controls to favor the hypothesis with stronger evidence. And when competing causal hypotheses were matched in terms of evidence but varied in their consistency with prior beliefs, children who were prompted to explain were significantly more likely than controls to favor the hypothesis with a higher prior. In other words, explaining made children more responsive to *both* crucial ingredients of causal learning: evidence and prior beliefs.

In their final study, Walker et al. considered a case in which evidence and prior beliefs came into conflict: a hypothesis that accounted for 100% of the evidence ("blue blocks activate the machine") was pitted against a hypothesis favored by prior beliefs ("big blocks activate the machine"), but that only accounted for 75% of the evidence. In this case, children who were prompted to explain were significantly more likely than controls to go with prior beliefs, guessing that a novel *big* block rather than a novel *blue* block would activate the machine. This pattern of responses was compared against the predictions of a Bayesian model that incorporated children's own priors and likelihoods as estimated from an independent task. The results suggested that children who were prompted to explain were *less* likely than children in the control condition to conform to Bayesian inference. This result may seem surprising in light of explainers' greater sensitivity to both evidence and prior beliefs, which suggests that explaining results in "better" performance. However, it's less surprising in light of the findings reported in the previous section, which consistently point to a divergence between explanation-based judgments and assessments of posterior probability.

While the evidence summarized thus far is restricted to preschool-aged children, it's likely that similar processes operate in older children and adults. For instance, Kuhn and Katz (2009) had fourth-grade children engage in a causal learning task that involved identifying the causes of earthquakes by observing evidence. The children subsequently participated in a structurally similar causal learning task involving an ocean voyage, where half were instructed to *explain* the basis for each prediction that they made, and those in a control group were not. When the same students completed the earthquake task in a post-test, those who had explained generated a smaller number of evidence-based inferences; instead, they seemed to rely more heavily on their (mistaken) prior beliefs, in line with the findings from Walker et al. (2015). In a classic study with 8th-grade students, Chi, De Leeuw, Chiu, and LaVancher (1994) prompted students to "self-explain" as they read a passage about the circulatory system, with students in the control condition instead prompted to read the text twice. Students who explained were significantly more likely to acquire an accurate causal model of the circulatory system, in part, they suggest, because explaining "involved the integration of new information into existing knowledge" – that is, the coordination of evidence with prior beliefs. Finally, evidence with adults investigating the effects of explanation in categorization tasks mirror the findings from Walker et al. (2015), with participants who explain both more responsive to evidence (Williams & Lombrozo, 2010) and more likely to recruit prior beliefs (Williams & Lombrozo, 2013).

Why does the process of explaining affect causal learning? One possibility is that explaining simply leads to greater attention or engagement. This is unlikely for a variety of reasons. Prior work has found that while explaining leads to some improvements in performance, it also generates systematic impairments. In one study, children prompted to explain were significantly less likely than controls to remember the color of a gear in a gear toy (Legare & Lombrozo, 2014); in another, they were significantly less likely to remember which sticker was placed on a block (Walker et al., 2014). Research with adults has also found that a prompt to explain can slow learning and increase error rates in a category learning task (Williams, Lombrozo, & Rehder, 2013). Moreover, the findings from the final study of Walker et al. (2015) suggest that prompting children to explain makes them look less, not more, like ideal Bayesian learners. Far from generating a global boost in performance, explanation seems to generate highly selective benefits.

A second possibility is that explaining plays a motivational role that's specifically tied to causal learning. In a provocatively-titled paper ("Explanation as orgasm and the drive for causal understanding"), Gopnik (1998, 2000) argues that the phenomenological satisfaction that accompanies a good explanation is part of what motivates us to learn about the causal structure of the world. Prompting learners to explain could potentially ramp up this motivational process, directing children and adults to causal relationships over causally-irrelevant details (consistent with Legare & Lombrozo, 2014; Walker et al., 2014). Explaining could also affect the course of causal inquiry itself, with effects on which data are acquired and how they inform beliefs (see Legare, 2012, for preliminary evidence that explanation guides exploration).

Finally (and not mutually exclusively), it could be that effects of explanation on learning are effectively a consequence of IBE – that is, that in the course of explaining, children generate explanatory hypotheses, and those explanatory hypotheses are evaluated with “loveliness” as a proxy for “likeliness.” For instance, in Walker et al. (2015), children may have favored the hypothesis that accounted for more evidence because it had greater scope or coverage, and the hypothesis consistent with prior knowledge because it provided a specification of mechanism or greater coherence. We suspect that this is mostly, but only mostly, correct. Some studies have found that children who are prompted to explain outperform those in control conditions *even when they fail to generate the right explanation,* or any explanation at all (Walker et al., 2014). This suggests the existence of some effects of *engaging in explanation* that aren’t entirely reducible to the effects of having generated any *particular* explanation.

While such findings are puzzling on a classic interpretation of IBE, they can potentially be accommodated with a modified and augmented version (Lombrozo, 2012; Wilkenfeld & Lombrozo, 2015). Wilkenfeld and Lombrozo (2015) argue for what they call “Explaining for the Best Inference,” an inferential practice that differs from IBE in focusing on the process of explaining as opposed to candidate explanations themselves. While IBE and EBI are likely to go hand in hand, there could be cases in which the explanatory processes that generate the best inferences aren’t identical with those promoted by possessing the best explanations, and EBI allows for this possibility.

In sum, there’s good evidence that the process of engaging in explanation influences causal learning. This is potentially driven by effects of explanation on the evaluation of *both* evidence and prior beliefs (Walker et al., 2015). One possibility is that by engaging in explanation, learners are more likely to favor hypotheses that offer “lovely” explanations (Lombrozo, 2012), and to engage in cognitive processes that affect learning even when a lovely or accurate explanation isn’t acquired (Wilkenfeld & Lombrozo, 2015). It’s not entirely clear, however, whether and when these effects of explanation lead to "better" causal learning. The findings from Amsterlaw & Wellman (2006) and Chi et al. (1994) suggest that effects can be positive, accelerating conceptual development and learning. Other findings are more mixed (e.g., Kuhn & Katz, 2009), with the modeling result from Walker et al. (2015) suggesting that prompting children to explain makes them integrate evidence and prior beliefs in a manner that corresponds less closely to Bayesian inference. Better delineating the contours of explanation's beneficial and detrimental effects will be an important step for future research. It will also be important to investigate how people's tendency to engage in explanation spontaneously corresponds to these effects. That is, are the conditions under which explaining is beneficial also the conditions under which people tend to spontaneously explain?

Assigning Causal Responsibility

The previous sections considered two kinds of causal reasoning, one involving novel causal structures and the other causal events generated by known structures. Another important class of causal judgments involves the assignment of *causal responsibility*: to which cause(s) do we attribute a given effect? For instance, a doctor might attribute her patient's disease to his weak immune system or to a cold virus, when both are in fact present and play a causal role.

Causal attribution has received a great deal of attention within social psychology, with the classic conundrum concerning the attribution of some behavior to a person ("she's so clumsy!") versus a situation ("the staircase is so slippery!") (see Fiske & Taylor, 2013, Kelley & Michela, 1980, and Malle, 2004, for reviews). While this research is often framed in terms of causation, it’s natural to regard attribution in terms of explanation, with attributions corresponding to an answer to the question of why some event occurred (“Why did Ava slip?”). In his classic “ANOVA model,” Kelley (1967, 1973) proposed that people effectively carry out an analysis of covariation between the behavior and a number of internal and external factors, such as the person, stimulus, and situation. For example, to explain why Ava slipped on the staircase yesterday, one would consider how this behavior fares along the dimensions of consensus (did other people slip?), the distinctiveness of the stimulus (did she slip only on that staircase?), and consistency across situations (does she usually slip, or was it the only time she did so?). Subsequent work, however, has identified a variety of additional factors that influence people’s attributions (e.g., Ahn, Kalish, Medin, & Gelman, 1995; Försterling, 1992; Hewstone & Jaspars, 1987; McArthur, 1972), and some have challenged the basic dichotomy on which the person-versus-situation analysis is based (Malle, 1999, 2004; Malle, Knobe, O’Laughlin, Pearce, & Nelson, 2000). We direct readers interested in social attribution to the chapter by Hilton (this volume).

Assignments of causal responsibility also arise in the context of what’s sometimes called “causal selection”: the problem of deciding which cause or causes in a chain or other causal structure best explain or account for some effect. Such judgments are especially relevant in moral and legal contexts, where they are closely tied to attributions of blame. For example, suppose that someone steps on a log, which pushes a boulder onto a picnic blanket, crushing a chocolate pie. The person, the log, and the boulder all played a causal role in the pie’s destruction, but various factors might influence our assignment of causal responsibility, including the location of each factor in the chain, whether and by how much it increased the probability of the outcome, and whether the person intended and foresaw the culinary catastrophe (see, e.g., Hart & Honoré, 1985; Hilton, McClure, & Sutton, 2009; Lagnado & Channon, 2008; McClure, Hilton, & Sutton, 2007; Spellman, 1997). The chapter by Lagnado and Gerstenberg on moral and legal reasoning (this volume) explores these issues in detail; also relevant is the chapter by Danks on singular causation (this volume).

While research has not (to our knowledge) investigated whether explanatory considerations such as simplicity and explanatory power influence judgments of causal responsibility, ideas from the philosophy and psychology of explanation can usefully inform research on this topic. For example, scholars of explanation often emphasize the ways in which an explanation request is underspecified by a why-question itself. When we ask "why did Ava slip on the stairs?", the appropriate response is quite different if we're trying to get at why *Ava* slipped (as opposed to Boris) than if we're trying to get at why Ava slipped *on the stairs* (as opposed to the landing). These questions involve a shift in what van Fraassen (1980) calls a "contrast class," i.e. the set of alternatives to the target event that the explanation should differentiate from the target via some appropriate relation (see also Cheng & Novick, 1991).

McGill (1989) showed in a series of studies that a number of previously established effects in causal attribution – effects of perspective (actor vs. observer; Jones & Nisbett, 1971), covariation information (consensus and distinctiveness; Kelley, 1967), and the valence of the behavior being explained (positive vs. negative; Weiner, 1985) – are related to shifts in the contrast class. Specifically, by manipulating the contrast class adopted by participants, McGill was able to eliminate the actor-observer asymmetry, interfere with the roles of consensus and distinctiveness information, and counteract self-serving attributions of positive versus negative performance. These findings underscore the close relationship between attribution and explanation.

Focusing on explanation is also helpful in bringing to the foreground questions of causal *relevance* as distinct from *probability*. In a 1996 paper, Hilton presented a set of studies designed to clearly differentiate these notions. In one study, Hilton showed that contextual information can influence the perceived "goodness" and relevance of an explanation without necessarily affecting its probability. For example, participants were asked to rate the following explanation of why a watch-face broke (an example adapted from Einhorn & Hogarth, 1986): "the watch broke because the hammer hit it." This explanation was rated as fairly good, relevant, and likely to be true; however, after learning that the hammer hit the watch during a routine testing procedure at a watch factory, participants' ratings of explanation quality and relevance dropped. In contrast, ratings of probability remained high, suggesting that causal relevance and the probability of an explanation can diverge, and that these two factors differ in their susceptibility to this contextual manipulation. It's possible that these effects were generated by a shift in contrast, from "why did this watch break now (as opposed to not breaking now)?" to "why did this watch break (as opposed to some other watch breaking)?"

More recently, Chin-Parker & Bradner (2010) showed that effects of background knowledge and implicit contrasts extend to the generation of explanations. They manipulated participants' background assumptions by presenting a sequence of causal events that either did or did not seem to unfold towards a particular functional outcome (when it did, the sequence appeared to represent a closed-loop system functioning in a self-sustaining manner). Participants' explanations of an ambiguous observation at the end of the sequence tended to invoke a failure of a system to perform its function in the former case, but featured proximal causes in the latter case. (In contrast to prior research, context did not affect explanation *evaluation* in this design).

Taken together, these studies offer another set of examples of how explanatory considerations (in this case, the contextually-determined contrast class) can influence causal judgments, and suggest that ascriptions of causal responsibility may vary depending on how they are framed: in terms of causal relevance and explanation, or in terms of probability and truth. It's also possible that considerations such as simplicity and scope play a role in assigning causal responsibility, above and beyond their roles in causal inference and learning. These are interesting questions for future research.

The Varieties of Causal Explanation

There's no agreed-upon taxonomy for explanations; in fact, even the distinction between causal and non-causal explanation generates contested cases. For instance, consider an example from Putnam (1975). A rigid board has a round hole and a square hole. A peg with a square cross-section passes through the square hole, but not the round hole. Why? Putnam suggests that this can be explained by appeal to the geometry of the rigid objects (which is not causal), without appeal to lower-level physical phenomena (which are presumably causal). Is this a case of non-causal explanation? It depends on whom you ask.

One taxonomy that has proven especially fruitful in the psychological study of explanation has roots in Aristotle's four causes (efficient, material, final, and formal), which are sometimes characterized not as causes per se, but in terms of explanation – as distinct answers to a "why?" question (Falcon, 2015). Efficient causes, which identify "the primary source of the change or rest" (e.g., a carpenter who makes a table), seem like the most canonically causal. Material causes, which specify "that out of which" something is made (e.g., wood for a table), are not causal in a narrow sense (for instance, we wouldn't say that the wood *causes* or is *a cause of* the table), but they nonetheless play a clear causal role in the production of an object. Final and formal causes are less clearly causal; but, as we consider below, there are ways in which each could be understood causally, as well.

First, consider final causes, which offer "that for the sake of which a thing is done." Final cause explanations (or perhaps more accurately, their contemporary counterparts) are also known as *teleological* or *functional* explanations, as they offer a goal or a function. For instance, we might explain the detour to the café by appeal to a goal (getting coffee), or the blade's sharpness by appeal to its function (slicing vegetables). On the face of it, these explanations defy the direction of causal influence: they explain a current event (the detour) or property (the sharpness) by appeal to something that occurs only *later* (the coffee-acquisition or the vegetable-slicing). Nonetheless, some philosophers have argued that teleological explanations can be understood causally (e.g., Wright, 1976), and there's evidence that adults (Lombrozo & Carey, 2006) and children (Kelemen & DiYanni, 2005) treat them causally, as well (see also Chaigneau, Barsalou, & Sloman, 2004, and Lombrozo & Rehder, 2012 for more general investigations of the causal structure of functions).

How can teleological explanations be causal? On Wright's view, teleological explanations don't explain the present by appeal to the future – rather, the appeal to an unrealized goal or function is a kind of shorthand for a complex causal process that brought about (and hence *preceded*) what's being explained. In cases of intentional action, the function or goal could be a shorthand for the corresponding *intention* that came first: the detour to the café was caused by a preceding intention to get coffee, and the blade's sharpness was caused by the designer's antecedent intention to create a tool for vegetable-slicing. Other cases, however, can be more complex. For instance, we might explain *this* zebra's stripes by appeal to their biological function (camouflage) because its ancestors had stripes that produced effective camouflage, and in part for that reason, stripes were increased or maintained in the population. If past zebra stripes didn't produce camouflage, then this zebra wouldn't have stripes (indeed, this zebra might not exist at all). In this case, the function can be explanatory because it was produced by "a causal process sensitive to the consequences of changes it produces" (Lombrozo & Carey, 2006; Wright, 1976), even in the absence of a preceding intention to realize the function.

Lombrozo and Carey (2006) tested these ideas as a descriptive account of the conditions under which adults accept teleological explanations. In one study, they presented participants with causal stories in which a functional property did or did not satisfy Wright's conditions. For example, participants learned about genetically-engineered gophers that eat weeds, and whose pointy claws damage the roots of weeds as they dig, making them popular among farmers. The causal role of "damaging roots" in bringing about the pointy claws varied across conditions, from no role (the genetic engineer *accidentally* introduced a gene sequence that resulted in gophers with pointy claws), to a causal role stemming from an intention to damage roots (the genetic engineer intended to help eliminate weeds, and to that end engineered pointy claws), to a causal role *without* an intention to damage roots (the genetic engineer didn't realize that pointy claws damaged weed roots, but did notice that the pointy claws were popular and decided to create all of his gophers with pointy claws). Participants then rated the acceptability and quality of teleological (and other) explanations. For the vignette involving genetically-engineered gophers, they were asked why the gophers had pointy claws, and rated "Because the pointy claws damage weed roots" as a response.

In this and subsequent studies, Lombrozo and Carey (2006) found that teleological explanations *are* understood causally in the sense that participants only accepted teleological explanations when the function or goal invoked in the explanation played an appropriate causal role in bringing about what was being explained. More precisely, this causal requirement was *necessary* for teleological explanations to be accepted, but not *sufficient*. In the examples above, teleological explanations were accepted at high levels when the function was intended, at moderate levels when the function played a non-intentional causal role, and at low levels when the function played no causal role at all. Lombrozo and Carey suggest (and provide evidence) that in addition to satisfying certain causal requirements, teleological explanations might call for the existence of a general pattern that makes the function predictively useful.

Kelemen and DiYanni (2005) conducted a study with elementary school children (6-7 and 9-10 year old) investigating the relationship between their acceptance and generation of teleological explanations for natural phenomena on the one hand, and their causal commitments concerning their origins on the other hand – specifically, whether they believed that an intentional designer of some kind ("someone or something") made them or they "just happened." The tendency to endorse and generate teleological explanations of natural events, non-living natural objects, and animals was significantly correlated with belief in the existence of an intentional creator of some kind, be it God, a human, or an unspecified force or agent. While these findings don't provide direct support for the idea that teleological explanations are grounded in a preceding intention to produce the specific function in question, the link between teleological explanations and intentional design more generally is consistent with the idea that teleological explanations involve some basic causal commitments. Along the same lines, Kelemen, Rottman, and Seston (2013) found that adults (including professional scientists) who believe in God or "Gaia" are more likely to accept scientifically-unwarranted teleological explanations (see also ojalehto, Waxman, & Medin, 2013, for a relevant discussion). Thus, the findings to date suggest that teleological explanations are understood causally by both adults and children.

What about formal explanations? Within Aristotle's framework, a formal explanation offers "the form" of something or "the account of what-it-is-to-be." Within psychology, what little work there is on formal explanation has focused on explanations that appeal to category membership. For example, Prasada and Dillingham (2006) define formal explanations as stating that tokens of a type have certain properties because they are the kinds of things they are (i.e. tokens of the respective type): we can say that Zach diagnoses ailments *because he is a doctor*, or that a particular object is sharp *because it is a knife*.

In their original paper and in subsequent work, Prasada and Dillingham (2006, 2009) argue that formal explanations are *not* causal, but instead explanatory by virtue of a part-whole relationship. They show that only properties that are considered to be *aspects* of the kind support formal explanations, in contrast to "statistical" properties that are merely reliably associated with the kind. For example, people accepted a formal explanation of why something has four legs by reference to its category ("because it's a dog"), and also accepted the claim that "having four legs" is one aspect of being a dog. In contrast, participants rejected formal explanations such as “that (pointing to a barn) is red because it’s a barn,” and also denied that being red is one aspect of being a barn (even though most barns are red). Prasada and Dillingham (2009) argue that the relationship underlying such formal explanation is constitutive (not causal): aspects are connected to kinds via a part-whole relationship, and such relationships are explanatory because the “existence of a whole presupposes the existence of its parts, and thus the existence of a part is rendered intelligible by identifying the whole of which it is a part” (p. 421).

Prasada and Dillingham offer two additional pieces of evidence for the proposal that formal explanations are constitutive, and not causal. First, they demonstrate the explanatory potential of the part-whole relationship by showing that when this relationship is made explicit, even statistical features can support formal explanations. For example, we can explain: “Why is that (pointing to a barn) red? Because it is a red barn,” where being red is understood as part of being a red barn (Prasada & Dillingham, 2009). This explanation isn’t great, but neither is it tautological: it identifies the source of the redness in something about the red barn as opposed, for instance, to the light that happens to be shining on it (see also Cimpian & Salomon, 2014, on “inherent” explanations). Less convincingly, they attempt to differentiate formal explanations from causal-essentialist explanations. On causal-essentialist accounts, a category’s essence is viewed as the cause of the category members’ properties (Gelman, 2003; Gelman & Hirschfeld, 1999; Medin & Ortony, 1989), which could ground formal explanations in a causal relationship. To test this, Prasada and Dillingham had participants evaluate explanations such as: “Why does that (pointing to a dog) have four legs? Because it has the essence of a dog which causes it to have four legs” (Prasada & Dillingham, 2006). While there was a trend for formal explanations to be rated more highly than causal-essentialist explanations for properties that were taken to be aspects of a given kind, the results were inconclusive. As Prasada and Dillingham acknowledge, the wording of the causal-essentialist explanations was awkward, which could partially account for their middling acceptance. It thus remains a possibility that at least some formal explanations are understood causally, as pointers to some category-associated essence or causal factor responsible for the properties being explained.

One reason it's valuable to recognize the diversity of explanations is because different kinds of explanations lead to systematically different patterns of causal judgment. For example, Lombrozo (2009) investigated the relationship between different kinds of causal explanations and the relative importance of features in classification (see also Ahn, 1998). Participants learned about novel artifacts and organisms with three causally-related features. To illustrate: one item involved "holings," a type of flower with brom compounds in its stem, which makes it bend over as it grows, which means its pollen can be spread to other flowers by wandering field mice. Participants were asked a why-question about the middle feature (e.g., "Why do holings typically bend over?"), which was ambiguous as a request for a mechanistic explanation (e.g., "Because of the brom compounds") or a teleological explanation (e.g., "In order to spread their pollen"). Participants provided an explanation and were subsequently asked to decide whether novel flowers were holings, where some shared the mechanistic feature (brom compounds) and some shared the functional feature (bending over). Lombrozo found that participants who provided functional explanations in response to the ambiguous why-question were significantly more likely than those who did not to privilege the functional feature relative to the mechanistic feature when it came to classification. Similarly, a follow-up study found that experimentally prompting participants to generate a particular explanation type by disambiguating the why-question ("In other words, what purpose might bending over serve?") had the same effect (see also Lombrozo & Rehder, 2012 for additional evidence about the relationship between functions and kind classification).

Additional studies suggest that effects of mechanistic versus functional explanations extend beyond judgments of category membership. Lombrozo and Gwynne (2014) employed a method similar to Lombrozo (2009), presenting participants with causal chains consisting of three elements, such as a certain gene that causes a speckled pattern in a plant, which attracts butterflies that play a role in pollination. Participants explained the middle feature (the speckled pattern) and generalized a number of aspects of that feature (e.g., its density, contrast, and color) to novel entities that shared either a causal or a functional feature with the original. Lombrozo and Gwynne found that explaining a property functionally (versus mechanistically) promoted the corresponding type of generalization.

Vasilyeva and Coley (2013) demonstrated a similar link between explanation and generalization in an open-ended task. Participants learned about plants and animals possessing novel but informative properties (e.g., *ducks* have *parasite X* [or *X-cells*]) and generated hypotheses about which other organisms might share the property. In the course of generating these hypotheses, participants spontaneously produced formal, causal, and teleological explanations in a manner consistent with the property they reasoned about. Most importantly, the type of explanation predicted the type of generalization: for example, people were most likely to generalize properties to entities related via causal interactions (e.g., plants and insects that ducks eat, or things that eat ducks) after generating causal explanations (e.g., they got it from their food). In a separate set of studies, Vasilyeva and Coley (in prep.) ruled out an alternative account based exclusively on direct effects of generalized properties on generalizations.

Beyond highlighting some causal relationships over others, different kinds of explanations could change the way participants represent and reason about causal structure. Indeed, findings from Lombrozo (2010) suggest this is the case. In a series of studies, Lombrozo presented participants with causal structures drawn from the philosophical literature and intended to disambiguate two accounts of causation: those based on some kind of *dependence* relationship (see Le Pelley, Over, this volume) and those based on some kind of *transference* (see Wolff, this volume). According to one version of the former view, C is a cause of E if it's the case that had C not occurred, E would not have occurred. In other words, E *depends* upon C in the appropriate way, in this case counterfactually. According to one version of transference views, C is a cause of E if there was a physical connection between C and E – some continuous mechanism or conserved physical quantity, such as momentum.

While dependence and transference often go hand in hand, they can come apart in cases of "double prevention" and "overdetermination." Lombrozo presented participants with such cases and found that judgments were more closely aligned with dependence views than transference views when the causal structures were directed towards a function or goal, and therefore supported a teleological explanation. Lombrozo (2010) explains this result, in part, by appeal to the idea of equifinality: when a process is goal-directed, the end may be achieved despite variations in the means. To borrow Williams James's famous example, Romeo will find his way to Juliet whatever obstacle is placed in his path (James, 1890). He might scale a fence or wade through a river, but the end –

reaching Juliet – will remain the same. When participants reason about a structure in teleological or goal-directed terms, they may similarly represent it as means- or mechanism-invariant, and therefore focus on dependence relationships irrespective of the specific transference that happened to obtain.

In sum, pluralism has long been recognized as a feature of explanation, with Aristotle's taxonomy providing a useful starting point for charting variation in explanations (although it is by no means the only taxonomy of explanation; see, for example, Cimpian & Salomon, 2014, on inherent versus extrinsic explanations). We've reviewed evidence that teleological explanations are causal explanations, but that they are nonetheless treated differently from mechanistic explanations, which do not appeal to functions or goals. The evidence concerning formal explanations is less conclusive, but points to a viable alternative to a causal interpretation, with formal explanation instead depending on constitutive "part-whole" relations.

One reason it's valuable to recognize explanatory pluralism is because it could provide a useful roadmap for thinking about pluralism when it comes to causation and causal relations. In fact, as we've seen, different kinds of explanations do lead to systematic differences in classification and inference, with evidence that causal relationships themselves may be represented differently under different "explanatory modes." In the following section, we take a closer look at mechanistic explanations and their relationship to causation and mechanisms.

Explanation and Causal Mechanisms

The "mechanistic explanations" considered in the previous section concerned the identification of one or more causes that preceded some effect. Often, however, causal explanations don't simply identify causes, but instead aim to articulate *how* the cause brought about the effect. That is, they involve a *mechanism*. But what, precisely, *is* a mechanism? Are all mechanisms causal? And do mechanisms have a privileged relationship to explanation? In this section, we begin to address these questions about the relationship between mechanisms and explanations. For a more general discussion of mechanisms, we direct readers to the chapter on mechanisms by Johnson and Ahn (this volume).

Within psychology, there is growing interest in the role of mechanisms in causal reasoning. For example, Ahn, Kalish, Medin and Gelman (1995) found that people seek "mechanistic" information in causal attribution. Park and Sloman (2013) found that people's violations of the Markov assumption depended on their "mechanistic" beliefs about the underlying causal structure. Buehner and McGregor (2006) showed that beliefs about mechanism type moderate effects of temporal contiguity in causal judgments (see also Ahn & Bailenson, 1996; Buehner & May, 2004; Fugelsang & Thompson, 2000; Koslowski & Okagaki, 1986; Koslowski, Okagaki, Lorenz, & Umbach, 1989 ; for reviews see Ahn & Kalish, 2013; Johnson & Ahn, this volume; Koslowski, 1996, 2012; Koslowski & Masnik, 2010; Sloman & Lagnado, 2014; Waldmann & Hagmayer, 2013). Despite these frequent appeals to mechanisms and mechanistic information, however, there isn't an explicit and widely endorsed conception of "mechanism" on offer.

Most often, a mechanism is taken to spell out the intermediate steps between some cause and some effect. For example, Park and Sloman (2014) define a mechanism as "the set of causes, enablers, disablers, and preventers that are directly involved in producing an effect, along with information about how the effect comes about, including how it unfolds over time" (p. 807). Research that adopts a perspective along these lines often goes further in explicitly identifying such mechanisms *as explanations* (and these terms are often used interchangeably, as in Koslowski & Masnik, 2010). Other work operationalizes mechanisms using measures of explanation, implicitly suggesting a correspondence. For example, to validate a manipulation of mechanism, Park and Sloman asked participants whether the same *explanation* applies to both effects in a common-cause structure (see also Park & Sloman, 2013). Similarly, in a study examining mental representations of mechanisms, Johnson and Ahn (2015) considered (but did not ultimately endorse) an "explanatory" sense of mechanism, which they operationalized by asking participants to rate the extent to which some event B *explains* why event A led to event C.

Shifting from psychology to philosophy, we find a class of accounts of explanation that likewise associate explanations with a specification of mechanisms (e.g., Bechtel & Abrahamsen, 2010; Glennan, 1996, 2002; Machamer, Darden, & Craver, 2000; Railton, 1978; Salmon, 1984). Consistent with the empirical work reviewed above, some of these accounts (e.g., Railton, 1978; Salmon, 1984) consider mechanisms to be "sequences of interconnected events" (Glennan, 2002, p. S345). Canonical examples include causal chains or networks of events leading to a specific outcome, such as a person who kicks a ball, which bounces off a pole, which breaks a window. On these views, explanation, causation, and mechanisms are not only intimately related, but potentially interdefined.

A second view of mechanisms within philosophy, however, departs more dramatically from work in psychology, and also suggests a more circumscribed role for causation. These views analyze mechanisms as complex systems that involve a (typically hierarchical) structure and arrangement of parts and processes, such as that exhibited by a watch, a cell, or a socioeconomic system (e.g., Bechtel & Abrahamsen, 2010; Bechtel & Richardson, 1993; Glennan, 1996, 2002; Machamer, Darden, & Craver, 2000). Within this framework, Craver and Bechtel (2007) offer an insightful analysis of causal *and non-causal* relationships within a multi-level mechanistic system. Specifically, they suggest that interlevel (i.e., "vertical") relationships within a mechanism aren't causal, but *constitutive*. For instance, a change in rhodopsin in retinal cells can partially explain how signal transduction occurs, but we wouldn't say that this change *causes* signal transduction; it arguably *is* signal transduction (or one aspect of it). Craver and Bechtel point out that constitutive relations conflict with many common assumptions about event causation: that causes and effects must be distinct events, that causes precede their effects, that the causal relation is asymmetrical, and so on. Unlike causation, explanation can accommodate both causal (intralevel) relationships and constitutive (interlevel) relationships, of the kind documented by Prasada and Dillingham's (2009) work on formal explanation.

Although Craver and Bechtel convincingly argue that the causal reading of interlevel relationships is erroneous (see also Glennan, 2010, for related claims), as a descriptive matter, it could be that laypeople nonetheless interpret them in causal terms. An example from the *Betty Crocker Cookbook,* discussed by Patricia Churchland (1994), illustrates the temptation. In the book, Crocker is correct to explain that microwave ovens work by accelerating the molecules comprising the food, but she wrongly states that the excited molecules rub against one another and that their friction *generates* heat. Crocker assumes that the increase in mean kinetic energy of the molecules *causes* heat, when in fact heat is *constituted* by the mean kinetic energy of the molecules (Craver & Bechtel, 2007). A study by Chi, Roscoe, Slotta, Roy and Chase (2012) showed that eighth and ninth graders, like Crocker, tended to misconstrue non-sequential, emergent processes as direct sequential causal relationships. It's possible that adults might make similar errors as well, assimilating non-causal explanations to a causal mold.

There are thus many open questions about how best to define mechanisms for the purposes of psychological theory, and about the extent to which mechanisms are represented in terms of strictly causal relationships. What we do know, however, is that explanations and mechanisms seem to share a privileged relationship. More precisely, there's evidence that the association between mechanisms and explanation claims is closer than that between mechanisms and corresponding causal claims (Vasilyeva & Lombrozo, 2015).

The studies by Vasilyeva and Lombrozo (2015) used "minimal pairs": causal and explanatory claims that were matched as closely as possible. For example, participants read about a person, PK, who spent some time in the portrait section of a museum and made an optional donation to the museum. They were then asked to evaluate how good they found an *explanation* for the donation ("Why did PK make an optional donation to the museum? Because PK spent some time in the portrait section"), or how strongly they endorsed a *causal relationship* ("Do you think there exists a causal relationship between PK spending some time in a portrait section and PK making an optional donation to the museum?").

Vasilyeva and Lombrozo varied two factors across items and participants: the strength of covariation evidence between the candidate cause and effect, and knowledge of a mediating mechanism. In the museum example, some participants learned the speculative hypothesis that "being surrounded by many portraits (as opposed to other kinds of paintings) creates a sense that one is surrounded by watchful others. This reminds the person of their social obligations, which in turn encourages them to donate money to the public museum." Both explanation and causal judgments were affected by these manipulations of covariation and mechanism information. However, they were not affected equally: specifying a mechanism had a stronger effect on explanation ratings than on causal ratings, while the strength of covariation evidence had a stronger effect on causal ratings than on explanation ratings.

The findings from Vasilyeva and Lombrozo (2015) support a special relationship between explanations and mechanisms. They also challenge views that treat explanations as equivalent to identifying causal relationships, since matched explanation and causal claims were differentially sensitive to mechanisms and covariation. The findings thus raise the possibility that explanatory and causal judgments are tuned to support different cognitive functions. For example, explanation could be especially geared towards reliable and broad generalizations (Lombrozo & Carey, 2006), which can benefit from mechanistic information: when we understand the mechanism by which some cause generates some effect, we can more readily infer whether the same relationship will obtain across variations in circumstances. By learning the mechanism that mediates the relationship between visiting a portrait gallery and making an optional museum donation, for example, we're in a better position to predict whether visiting a figurative versus an abstract sculpture garden will have the same effect. This benefit can potentially be realized with quite skeletal mechanistic (Rozenblit & Keil, 2002) or functional understanding (Alter, Oppenheimer, & Zemla, 2010); people need not understand a mechanism in full detail to gain some inferential advantage. Causal claims, by contrast, could more closely track the evidence concerning a particular event or relationship, rather than the potential for broad generalization.

In sum, the picture that emerges is one of partial overlap between causality, explanation, and mechanisms. Work in philosophy offers a variety of proposals emphasizing different aspects of mechanisms: structure, functions, temporally unfolding processes connecting starting conditions to the end state, and so on. Explanatory and causal judgments could track different aspects of mechanisms, resulting in the patterns of association and divergence observed. We suspect that adopting more explicit and sophisticated notions of mechanism will help research in this area move forward. On a methodological note, we think the strategy adopted in Vasilyeva and Lombrozo (2015) – of contrasting the characteristics of causal explanation claims with "matched" causal claims – could be useful in driving a wedge between different kinds of judgments, thus shedding light on their unique characteristics and potentially unique roles in human cognition. This strategy can also generalize to other kinds of judgments. For example, Dehghani, Iliev, and Kaufmann (2012) and Rips and Edwards (2013) both report systematic patterns of divergence between explanations and counterfactual claims, another judgment with a potentially foundational relationship to both explanation and causation.

Concluding Remarks

Throughout the chapter, we've seen good evidence that explanatory considerations affect causal reasoning, with implications for causal inference, causal learning, and attribution. We've also considered different kinds of explanations, including their differential effects on causal generalizations and causal representation, and the role of mechanisms in causal explanation. However, many questions remain open. We highlight four especially pressing questions here.

First, we've observed many instances in which explanation leads to departures from "normative" reasoning, at least on the assumption that one ought to infer causes and causal relationships by favoring causal hypotheses with the highest posterior probabilities. Are these departures truly errors? Or have we mischaracterized the relevant competence? In particular, could it be that explanatory judgments are well-tuned to some cognitive end, but that end is not the approximation of posterior probabilities?

Second, we've focused on a characterization of explanations and the effects of engaging in explanation, with little attention to underlying cognitive mechanisms. How do people actually go about generating and evaluating causal explanations? How do the mental representations that support explanation relate to those that represent causal structure? And how do explanatory capacities arise over the course of development?

Third, what's the relationship between causal and non-causal explanations? Are they both explanatory by virtue of some shared explanatory relationship, or are causal explanations explanatory by virtue of being causal, with non-causal explanations explanatory for some other reason (for instance, because they embody a part-whole relationship)? On each view, what are the implications for causation?

Finally, we've seen how debates in explanation (from both philosophy and psychology) can inform the study of causation, with examples including inference to the best explanation, the idea of a "contrast class," and pluralism about explanatory kinds. Can the literature on *levels* of explanation (e.g., Potochnik, 2010) perhaps inspire some new debates about *levels* of causation (as in, e.g., Woodward, 2010)? Recent work on hierarchical

Bayesian models and hierarchical causal structures are beginning to move in this direction, with the promise of a richer and more powerful way to understand humans' remarkable ability to reason about and explain the causal structure of the world.

Acknowledgments

The preparation of this chapter was partially supported by the *Varieties of Understanding Project* funded by the Templeton Foundation, as well as an NSF CAREER award to the first author (DRL-1056712). We are also grateful to David Danks, Samuel Johnson, and Michael Waldmann for helpful comments on a previous draft if this chapter.

References

Ahn, W. (1998). Why are different features central for natural kinds and artifacts?: The role of causal status in determining feature centrality. *Cognition*, *69*(2), 135–178. doi:10.1016/S0010-0277(98)00063-8

Ahn, W. K., & Bailenson, J. (1996). Causal attribution as a search for underlying mechanisms: An explanation of the conjunction fallacy and the discounting principle. *Cognitive Psychology*, *31*(1), 82–123. doi:10.1006/cogp.1996.0013

Ahn, W. K., Kalish, C. W., Medin, D. L., & Gelman, S. a. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, *54*, 299–352. doi:10.1016/0010-0277(94)00640-7

Ahn, W. K., & Kalish, C. (2002). The role of mechanism beliefs in causal reasoning. In F. C. Keil (Ed.), *Explanation and cognition*. MIT Press.

Alter, A. L., Oppenheimer, D. M., & Zemla, J. C. (2010). Missing the trees for the forest: A construal level account of the illusion of explanatory depth. *Journal of Personality and Social Psychology*, *99*, 436–451. doi:10.1037/a0020218

Amsterlaw, J., & Wellman, H. M. (2006). Theories of mind in transition: A microgenetic study of the development of false belief understanding. *Journal of Cognition and Development*, *7*(2), 139–172. doi:10.1207/s15327647jcd0702_1

Baker, A. (2013). Simplicity. *In E.N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy.* (Fall 2013 Edition). URL retrieved May 2015 from <http://plato.stanford.edu/archives/fall2013/entries/simplicity/>

Bartsch, K., & Wellman, H. M. (1995). *Children talk about the mind*. Oxford university press.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*(1995), 421–441. doi:10.1016/j.shpsc.2005.03.010

Bonawitz, E. B., & Lombrozo, T. (2012). Occam’s rattle: Children's use of simplicity and probability to constrain inference. *Developmental Psychology*, *48*(4), 1156–1164. doi:10.1037/a0026471

Buehner, M. J. (2005). Contiguity and covariation in human causal inference. *Learning & Behavior : A Psychonomic Society Publication*, *33*(2), 230–238. doi:10.3758/BF03196065

Buehner, M. J., & May, J. (2004). Abolishing the effect of reinforcement delay on human causal learning. *The Quarterly Journal of Experimental Psychology*, *57B*, 179–191.

Buehner, M. J., & McGregor, S. (2006). Temporal delays can facilitate causal attribution: Towards a general timeframe bias in causal induction. *Thinking & Reasoning*, *12*, 353–378.

Chaigneau, S. E., Barsalou, L. W., & Sloman, S. A. (2004). Assessing the causal structure of function. *Journal of Experimental Psychology. General*, *133*(4), 601–25. doi:10.1037/0096-3445.133.4.601

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367–405.

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of personality and social psychology*, *58*(4), 545.

Cheng, P. W., & Novick, L. R. (1991). Causes versus enabling conditions. *Cognition*, *40*(1-2), 83–120. doi:10.1016/0010-0277(91)90047-8

Chi, M. T. H., De Leeuw, N., Chiu, M.-H., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*(3), 439–477. doi:10.1207/s15516709cog1803_3

Chi, M. T. H., Roscoe, R. D., Slotta, J. D., Roy, M., & Chase, C. C. (2012). Misconceived causal explanations for emergent processes. *Cognitive Science*, *36*(1), 1-61.

Churchland, P. S. (1994). Can neurobiology teach us anything about consciousness? *Proceedings and Addresses of the American Philosophical Association*, *67*(4), 23–40. doi:10.2307/3130741

Cimpian, A., & Salomon, E. (2014). The inherence heuristic: An intuitive means of making sense of the world, and a potential precursor to psychological essentialism. *Behavioral and Brain Sciences, 37*(5), 461–480.

Craver, C. F., & Bechtel, W. (2007). Top-down causation without top-down causes. *Biology and Philosophy*, *22*, 547–563. doi:10.1007/s10539-006-9028-8

Dehghani, M., Iliev, R., & Kaufmann, S. (2012). Causal explanation and fact mutability in counterfactual reasoning. *Mind & Language*, *27*(1), 55–85. doi:10.1111/j.1468-0017.2011.01435.x

Douven, I. (2011). Abduction. *In E.N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy.* (Spring 2011 Edition). URL retrieved May 2015 from <http://plato.stanford.edu/archives/spr2011/entries/abduction/>

Douven, I. (2013). Inference to the best explanation, Dutch books, and inaccuracy minimisation. *Philosophical Quarterly*, *63*(252), 428–444. doi:10.1111/1467-9213.12032

Douven, I., & Schupbach, J. N. (2015a). The role of explanatory considerations in updating. *Cognition*, *142,* 299-311. doi:10.1016/j.cognition.2015.04.017

Douven, I., & Schupbach, J. N. (2015b). Probabilistic alternatives to Bayesianism: the case of explanationism. *Frontiers in Psychology*, *6*, 1–9. doi:10.3389/fpsyg.2015.00459

Falcon, A. (2015). Aristotle on causality. In E.N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy* (Spring 2015 Edition). URL retrieved May 2015 from <http://plato.stanford.edu/archives/spr2015/entries/aristotle-causality/>

Fiske, S. T., & Taylor, S. E. (2013). *Social cognition: From brains to culture*. Sage.

Försterling, F. (1992). The Kelley model as an analysis of variance analogy: How far can it be taken? *Journal of Experimental Social Psychology*, *28*(5), 475–490. doi:10.1016/0022-1031(92)90042-I

Fugelsang, J. A., & Thompson, V. A. (2000). Strategy selection in causal reasoning: When beliefs and covariation collide. *Canadian Journal of Experimental Psychology*, *54*, 15–32.

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. Oxford University Press.

Gelman, S. A., & Hirschfeld, L. A. (1999). How biological is essentialism. *Folkbiology*, *9*, 403–446.

Glennan, S. (1996). Mechanisms and the nature of causation. *Erkenntnis*, *44*(1), 49–71.

Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of Science*, *69*(3), pp. S342–S353.

Glennan, S. (2010). Mechanisms, causes, and the layered model of the world. *Philosophy and Phenomenological Research*, *81*(2), 362–381. doi:10.1111/j.1933-1592.2010.00375.x

Glymour, C., & Cheng, P. W. (1998). Causal mechanism and probablity: A normative approach. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 295-313). Oxford, UK: Oxford University Press.

Good, I. J. (1960). Weight of evidence, corroboration, explanatory power, information and the utility of experiments. *Journal of the Royal Statistical Society. Series B (Methodological)*, 319–331.

Gopnik, A. (1998). Explanation as orgasm. *Minds and Machines*, *8*(1), 101–118. doi:10.1023/A:1008290415597

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. 299–323.

Gopnik, A., & Sobel, D. M. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development*, *71*(5), 1205–1222. doi:10.1111/1467-8624.00224

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–84. doi:10.1016/j.cogpsych.2005.05.004

Harman, G. H. (1965). The inference to the best explanation. *Philosophical Review*, *74*(1), 88-95.

Hart, H. L. A., & Honoré, T. (1985). *Causation in the Law*. Oxford University Press.

Henderson, L. (2013). Bayesianism and inference to the best explanation. *The British Journal for the Philosophy of Science*, 0, 1-29.

Hewstone, M., & Jaspars, J. (1987). Covariation and causal attribution: A Logical Model of the intuitive analysis of variance. *Journal of Personality and Social Psychology*, *53*(4), 663-672.

Hilton, D. J., McClure, J., & Sutton, R. M. (2009). Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes? *European Journal of Social Psychology*, *39*, 1–18.

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: the new synthesis. *Annual Review of Psychology*, *62*, 135–163. doi:10.1146/annurev.psych.121208.131634

Johnson, S. G. B., & Ahn, W. (2015). Causal networks or causal islands? The Rrepresentation of mechanisms and the transitivity of causal judgment. *Cognitive Science*, 1-36. doi:10.1111/cogs.12213

Johnson, S. G. B., Johnston, A. M., Toig, A. E., & Keil, F. C. (2014). Explanatory scope informs causal strength inferences. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 2453-2458). Austin, TX: Cognitive Science Society.

Johnston, A. M., Johnson, S. G. B., Koven M. L., & Keil, F. C. (2015). Probabilistic versus heuristic accounts of explanation in children: Evidence from a latent scope bias. In D.C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (p. 1021-1026). Austin, TX: Cognitive Science Society.

Kelemen, D., & DiYanni, C. (2005). Intuitions about origins: Purpose and intelligent design in children's reasoning about nature. *Journal of Cognition and Development, 6*, 3-31. doi:10.1207/s15327647jcd0601_2

Kelemen, D., Rottman, J., & Seston, R. (2012). Professional physical scientists display tenacious teleological tendencies: Purpose-based reasoning as a cognitive default. *Journal of Experimental Psychology: General*, *142*(4), 1074–1083. doi:10.1037/a0030399

Kelley, H. H. (1967). Attribution theory in social psychology. *Nebraska Symposium on Motivation*, *15*, 192–238.

Kelley, H. H. (1973). The process of causal attributions. *American Psychologist*, *28*, 107–128.

Kelley, H. H., & Michela, J. L. (1980). Attribution theory and research. *Annual Review of Psychology*, *31*, 457–501. doi:10.1146/annurev.ps.31.020180.002325

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

Koslowski, B. (2012). Scientific reasoning: Explanation, confirmation bias, and scientific practice. In G. Feist & M. Gorman (Eds.), *Handbook of the Psychology of Science*. New York: Springer Publishing.

Koslowski, B., & Masnick, A. (2010). Causal reasoning and explanation. In *The Wiley-Blackwell Handbook of Childhood Cognitive Development, Second edition* (pp. 377–398). doi:10.1002/9781444325485.ch14

Koslowski, B., & Okagaki, L. (1986). Non-Humean indices of causation in problem-solving situations: Causal mechanism, analogous effects, and the status of rival alternative accounts. *Child Development*, *57*(5), 1100–1108.

Koslowski, B., Okagaki, L., Lorenz, C., & Umbach, D. (1989). When covariation is not enough: The role of causal mechanism, sampling method, and sample size in causal reasoning. *Child Development*, *60*(6), 1316–1327. doi:10.2307/1130923

Kuhn, D., & Katz, J. (2009). Are self-explanations always beneficial? *Journal of Experimental Child Psychology*, *103*(3), 386–394.

Lagnado, D. a, & Channon, S. (2008). Judgments of cause and blame: the effects of intentionality and foreseeability. *Cognition*, *108*(3), 754–70. doi:10.1016/j.cognition.2008.06.009

Legare, C. H. (2012). Exploring explanation: explaining inconsistent evidence informs exploratory, hypothesis-testing behavior in young children. *Child Development*, *83*(1), 173–85. doi:10.1111/j.1467-8624.2011.01691.x

Legare, C. H., & Lombrozo, T. (2014). Selective effects of explanation on learning during early childhood. *Journal of Experimental Child Psychology*, *126*, 198–212. doi:10.1016/j.jecp.2014.03.001

Legare, C. H., Wellman, H. M., & Gelman, S. A. (2009). Evidence for an explanation advantage in naïve biological reasoning. *Cognitive Psychology*, *58*(2), 177–94. doi:10.1016/j.cogpsych.2008.06.002

Lipton, P. (2004). *Inference to the best explanation*. Psychology Press.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*(3), 232–257. doi:10.1016/j.cogpsych.2006.09.006

Lombrozo, T. (2009). Explanation and categorization: how “why?” informs “what?”. *Cognition*, *110*(2), 248–53. doi:10.1016/j.cognition.2008.10.007

Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303–32. doi:10.1016/j.cogpsych.2010.05.002

Lombrozo, T. (2012). Explanation and abductive inference. In *Oxford Handbook of Thinking and Reasoning* (pp. 260–276). Oxford University Press. doi:10.1093/oxfordhb/9780199734689.013.0014

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, *99*(2), 167–204. doi:10.1016/j.cognition.2004.12.009

Lombrozo, T., & Gwynne, N. Z. (2014). Explanation and inference: mechanistic and functional explanations guide property generalization. *Frontiers in Human Neuroscience*, *8*(September), 700. doi:10.3389/fnhum.2014.00700

Lombrozo, T., & Rehder, B. (2012). Functions in biological kind classification. *Cognitive Psychology*, *65*(4), 457–485. doi:10.1016/j.cogpsych.2012.06.002

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*(1), 1-25. doi:10.1086/392759

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

Malle, B. F. (1999). How people explain behavior: a new theoretical framework. *Personality and Social Psychology Review, 3*(1), 23–48. doi:10.1207/s15327957pspr0301_2

Malle, B. F., Knobe, J., O’Laughlin, M. J., Pearce, G. E., & Nelson, S. E. (2000). Conceptual structure and social functions of behavior explanations: Beyond person-situation attributions. *Journal of Personality and Social Psychology*, *79*(3), 309–326. doi:10.1037//0022-3514.79.3.309

McArthur, L. A. (1972). The how and what of why: Some determinants and consequences of causal attribution. *Journal of Personality and Social Psychology*, *22*(2), 171–193.

McClure, J., Hilton, D. J., & Sutton, R. M. (2007). Judgments of voluntary and physical causes in causal chains: Probabilistic and social functionalist criteria for attributions. *European Journal of Social Psychology*, *37*, 879–901.

McGill, A. L. (1989). Context effects in judgments of causation. *Journal of Personality and Social Psychology*, *57*(2), 189–200.

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. *Similarity and Analogical Reasoning*, 179–195.

Ojalehto, B., Waxman, S. R., & Medin, D. L. (2013). Teleological reasoning about nature: Intentional design or relational perspectives? *Trends in Cognitive Sciences*, *17*(4), 166–171. doi:10.1016/j.tics.2013.02.006

Pacer, M., & Lombrozo, T. (2015). Ockham's Razor cuts to the root: simplicity in causal explanation. Manuscript in revision.

Pacer, M., Williams, J., Chen, X., Lombrozo, T., & Griffiths, T. (2013). Evaluating computational models of explanation using human judgments. *arXiv Preprint arXiv:1309.6855*.

Park, J., & Sloman, S. A. (2014). Causal explanation in the face of contradiction. *Memory & Cognition*, *42*(5), 806–20. doi:10.3758/s13421-013-0389-3

Park, J., & Sloman, S. A. (2013). Mechanistic beliefs determine adherence to the markov property in causal reasoning. *Cognitive Psychology*, *67*(4), 186–216. doi:10.1016/j.cogpsych.2013.09.002

Peirce, C. S. (1955). Abduction and induction. In *Philosophical writings of Peirce* (Vol. 11). Dover, New York.

Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the Story Model for juror decision making. *Journal of Personality and Social Psychology*, *62*(2), 189–206. doi:10.1037/0022-3514.62.2.189

Perales, J. C., & Shanks, D. R. (2003). Normative and descriptive accounts of the influence of power and contingency on causal judgement. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *56*(6), 977–1007. doi:10.1080/02724980244000738

Pine, K. J., & Siegler, R. S. (2003). The role of explanatory activity in increasing the generality of thinking. *Paper presented at the biennial meeting of the Society for Research in Child Development,* Tampa, FL.

Prasada, S., & Dillingham, E. M. (2006). Principled and statistical connections in common sense conception. *Cognition*, *99*(1), 73–112. doi:10.1016/j.cognition.2005.01.003

Prasada, S., & Dillingham, E. M. (2009). Representation of principled connections: a window onto the formal aspect of common sense conception. *Cognitive Science*, *33*(3), 401–48. doi:10.1111/j.1551-6709.2009.01018.x

Preston, J., & Epley, N. (2005). Explanations versus applications: The explanatory power of valuable beliefs. *Psychological Science*, *16*(10), 826–832. doi:10.1111/j.1467-9280.2005.01621.x

Putnam, H. (1975). Philosophy and our mental life. In H. Putnam, *Mind, Language and Reality: Philosophical Papers* (Vol. 2). New York: Cambridge University Press.

Read, S. J., & Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, *65*(3), 429.

Rips, L. J., & Edwards, B. J. (2013). Inference and explanation in counterfactual reasoning. *Cognitive Science*, *37*(6), 1107–35. doi:10.1111/cogs.12024

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*, 521–562. doi:10.1016/S0364-0213(02)00078-2

Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton University Press.

Schupbach, J. N., & Sprenger, J. (2011). The logic of explanatory power*. *Philosophy of Science*, *78*(1), 105–127. doi:10.1086/658111

Shanks, D. R., & Dickinson, A. (1988). Associative accounts of causality judgment. *Psychology of Learning and Motivation - Advances in Research and Theory*, *21*(C), 229–261. doi:10.1016/S0079-7421(08)60030-4

Siegler, R. S. (1995). How does change occur: a microgenetic study of number conservation. *Cognitive Psychology*, *28*, 225–273. doi:10.1006/cogp.1995.1006

Sloman, S. A, & Lagnado, D. (2014). Causality in thought. *Annual Review of Psychology, 66*, 223-247. doi:10.1146/annurev-psych-010814-015135

Spellman, B. A. (1997). Crediting causality. *Journal of Experimental Psychology: General, 126*(4), 323-348. doi:10.1037/0096-3445.126.4.323

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, *12*, 435–502.

Vasilyeva, N., & Coley, J.C. (2013). Evaluating two mechanisms of flexible induction: Selective memory retrieval and evidence explanation. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*). Austin, TX: Cognitive Science Society.

Vasilyeva, N., & Lombrozo, T. (2015). Explanation and causal judgments are differentially sensitive to covariation and mechanism information. *Proceedings of the 37th Annual Conference of the Cognitive Science Society.*

Waldmann, M. R., & Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition*, *82*(1), 27–58. doi:10.1016/S0010-0277(01)00141-X

Waldmann, M. R., & Hagmayer, Y. (2013). Causal reasoning. In D. Reisberg (Ed.), *Oxford Handbook of Cognitive Psychology* (pp. 733–752). New York: Oxford University Press.

Walker, C. M., Lombrozo, T., Legare, C. H., & Gopnik, A. (2014). Explaining prompts children to privilege inductively rich properties. *Cognition*, *133*(2), 343–57. doi:10.1016/j.cognition.2014.07.008

Walker, C.M., Lombrozo, T., Williams, J.J., Rafferty, A., & Gopnik, A. (2015). Explaining constrains causal learning in childhood. Manuscript in revision.

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review*, *92*(4), 548-573.

Wellman, H. M. (2011). Reinvigorating explanations for the study of early cognitive development. *Child Development Perspectives*, *5*(1), 33–38. doi:10.1111/j.1750-8606.2010.00154.x

Wellman, H. M., & Lagattuta, K. H. (2004). Theory of mind for learning and teaching: The nature and role of explanation. *Cognitive Development*, *19*, 479–497. doi:10.1016/j.cogdev.2004.09.003

Wellman, H. M., & Liu, D. (2007). Causal reasoning as informed by the early development of explanations. In A. Gopnik & L. Schulz (Eds.), *Causal Learning: Psychology, Philosophy, and Computation*, pp. 261–279.

Wilkenfeld, D.A., & Lombrozo, T. (2015). Infernece to the Best Explanation (IBE) vs. Explaining for the Best Inference (EBI). Manuscript under review.

Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: evidence from category learning. *Cognitive Science*, *34*(5), 776–806. doi:10.1111/j.1551-6709.2010.01113.x

Williams, J. J., & Lombrozo, T. (2013). Explanation and prior knowledge interact to guide learning. *Cognitive Psychology*, *66*(1), 55–84. doi:10.1016/j.cogpsych.2012.09.002

Williams, J. J., Lombrozo, T., & Rehder, B. (2013). The hazards of explanation: overgeneralization in the face of exceptions. *Journal of Experimental Psychology. General*, *142*(4), 1006–14. doi:10.1037/a0030996

Woodward, J. (2010). Causation in biology: stability, specificity, and the choice of levels of explanation. *Biology & Philosophy*, *25*(3), 287–318.

Wright, L. (1976). *Teleological explanations: An etiological analysis of goals and functions*. Berkeley, CA: University of California Press.