Word count: 5,334 (6,905 including references)

Explanation[1]

Tania Lombrozo

Abstract: Explanation has been an important topic of study in philosophy of science, in epistemology, and in other areas of philosophy. In parallel, psychologists have been studying children's and adults' explanations, including their role in inference and in learning. This entry reviews recent work that begins to bridge the philosophy and psychology of explanation, with sections introducing recent empirical work on explanation by philosophers, formal and functional accounts of explanation, inference to the best explanation, the role of explanation in discovery, and the implications of empirical work on explanation for the "negative program" in experimental philosophy.

Keywords: explanation, experimental philosophy, abduction, inference to the best explanation, simplicity

[1] Thanks to L.A. Paul, Dillon Sterling Plunkett, Daniel Wilkenfeld, the editors of this volume, and an additional referee for valuable comments on this article. Preparation of this article was partially supported by NSF grant DRL-1056712, the McDonnell Foundation, and the Varieties of Understanding Project, funded by the Templeton Foundation.

I. Introduction

The philosophical study of explanation has had a vexed relationship with psychology. With the rise of logical positivism in the first half of the last century, some philosophers adopted a psychological but deflationary stance, suggesting that explanation is mere "reduction to the familiar" (e.g., see Dray 1964, as cited by Friedman 1974). Others adopted a deliberately *anti*-psychologistic stance, aiming to ground explanations in objective relations or features of the world (e.g., Hempel 1965; Salmon 1984). Yet other twentieth-century accounts incorporated substantive, non-deflationary roles for human psychology, for instance characterizing explanations in terms of the understanding they produce (Achinstein 1983; Schank 1986), or analyzing them as contextually-sensitive answers to particular kinds of questions (Bromberger 1966; van Frassen 1980).

More recent work has opened the door to new approaches to an empirically-informed philosophy of explanation. Contemporary philosophy of science, for example, has sometimes focused on explanation in the special sciences, with detailed analyses of the role of explanation for working scientists (e.g., Craver & Darden 2013; Boumans 2009; Weiskopf 2011). Formal epistemologists (and others) have developed formal accounts of explanation and explanatory virtues (see Schupbach, this volume), often tied to work in related disciplines – such as statistics and computer science – and with corresponding tests against human judgments (Schupbach, 2011). At the same time, empirical work on children's and adults' explanations is flourishing, with a growing literature supporting the idea that explanations play an important role in human reasoning and conceptual representation (Lombrozo 2012; Wellman 2011). These developments have been partially informed by philosophy (Lombrozo, 2012), and have also begun to inform contemporary accounts of explanation in philosophy of science (e.g., Douven 2011; Wilkenfeld 2014).

Despite these connections between the philosophy and psychology of explanation, relatively little work on explanation has proceeded under the banner of "experimental philosophy." Ironically, this may be precisely because explanation is so closely tied to human psychology: the connections have been appreciated, if not always endorsed or pursued, by both philosophers and psychologists well before the rise of this relatively new movement. Nonetheless, there have been a handful of papers on what might be called "the experimental philosophy of explanation," and a great many more that lay important groundwork for those hoping to develop psychologically-informed accounts of explanation. This research is the focus of the current entry.

Given the scope of work on explanation within both philosophy and psychology, a great deal of interesting and important work is neglected. For a more thorough review of recent developments in the psychology of explanation, readers are directed to Lombrozo (2012), and to Wellman (2011) for work in developmental psychology. For reviews of work on the philosophy of explanation, readers are directed to Salmon (1989) and Woodward (2011). Finally, this review also excludes relevant work on closely-related topics, such as causation (e.g., Hitchcock 2012; Schaffer 2014; Woodward 2013) and understanding (e.g., Kvanvig 2015). Readers are also encouraged to consult related entries in this volume on various aspects of causation (Danks, this volume; Park & Sloman, this volume; Livengood & Rose, this volume) and on experimental philosophy of science (Machery, this volume).

I begin by considering work on folk and scientific conceptions of explanation, followed by functional and formal analyses of explanation. I then consider the impact of explanatory considerations on cognition, in particular as a guide to inference and discovery. The final sections consider how research on explanation might inform the "negative program" in experimental philosophy and identify important directions for future research.

2. Folk and scientific conceptions of explanation

Recently, some philosophers have turned to both lay and expert scientific judgments concerning the application of the word "explanation" to better understand the scope and viability of accounts of explanation. For example, Overton (2012) uses text-mining techniques to analyze uses of the word "explain" in the journal *Science*. He finds that over 45 percent of the 781 articles analyzed contain the word "explain" (or a close variant, such as "explanation" or "explicate"), that over 90 percent include the word "because," and that both of these figures are considerably larger than those found for two non-scientific corpora, a sample of English literature drawn from Project Gutenberg and a set of news articles from the Reuters news agency. He also finds that uses of "explain" are often qualified (e.g., "could explain," "may not explain") and are most likely to appear at the very beginning or end of articles rather than the middle. These findings suggest that explanation is indeed an important part of scientific discourse, perhaps especially in motivating and interpreting empirical work.

In a second set of analyses, Overton (2012) sampled 25 uses of "explain," "explains" or "explained" from abstracts and articles and classified the forms of the proffered explanations based on the nature of the explanans and the explanandum – i.e., whether each involved data, entities, kinds, models, or theories. The most interesting result was that the explanans always belonged to the same category or a more general category than the explanandum. For example, explanations in which some quality of a model explained some quality of a kind were common, but he found no instances in which some quality of a kind explained some quality of a model. A subsequent analysis identified many cases of "inference to the best explanation," which illustrated the reverse pattern: researchers used a more specific category, such as "data" or "entities," to draw inferences about a category that was at least as general, such as "kinds" or "theories." These final analyses were based on small samples and a coding scheme that not all researchers are likely to endorse, but they illustrate the method's potential as a way to learn about explanatory practices in science and about scientists' own conceptions of explanation.

Waskan et al. (2014) used an experimental manipulation to investigate whether the actual or potential achievement of intelligibility or understanding is central to scientists' and laypeople's concept of explanation. In an initial experiment, on-line participants were presented with a vignette in which a scientist offered a model of how gamma-ray bursts are produced in distant galaxies. The critical manipulation was whether the model was described as being *actually intelligible* to a normal human, *potentially intelligible* to a normal human, or *never intelligible*, due to its overwhelming complexity. The researchers found that participants were significantly more wiling to say that the model constituted an explanation when it was *actually intelligible* than when it was either *potentially* or *never intelligible*, suggesting that actual intelligibility could be a necessary condition for explanation, at least according to non-scientists. More generally, the findings suggest that psychological states, and not just nomic or other objective relations, play a role in the folk notion of explanation.

A subsequent experiment employed a more indirect method for assessing folk and scientific conceptions of explanation. In the experiment, participants read a modified popular science article about a complex model of gamma-ray bursts developed by "Dr. Brown." They were later asked whether the text had contained the following sentence, which never actually appeared: "Dr. Brown's paper and the accompanying computer model provide an explanation for why type-B2 stars produce gamma-ray bursts." The researchers reasoned that participants might misremember having read this sentence if they believed that Dr. Brown's contribution constituted an explanation. Again, features of the model's intelligibility were varied across conditions. In one case, Dr. Brown was eventually able "to detect some high-level (coarse-grained) structures and behavioral patterns that enabled him to make sense of why each distinct new simulation gravitated towards the same end state… Not even a mediocre astrophysicist could, after studying this report and its accompanying materials, fail to understand why type-B2 stars produce gamma-ray bursts." In the *potentially intelligible* condition, participants were told that no one had yet detected relevant high-level variables from Dr. Brown's materials, but that it was only a matter of time before someone did. Finally, the *never intelligible* condition indicated that due to the model's complexity, no one would ever be able to use it to understand the origin of gamma-ray bursts.

In this second experiment, over 80% of participants misremembered reading the explanation statement in the *intelligible* condition, which was significantly more often than the roughly 50% error rate observed in both the *potentially* and *never intelligible* conditions (which did not differ from each other). A replication with practicing scientists as participants yielded the same pattern of results, although the overall error rates were lower. As in the more explicit task used in their first study, these memory findings from

Waskan et al. suggest that bringing about a particular psychological state – namely some form of intelligibility or understanding – is important to explanation, and that the understanding must be *achieved*, not merely *achievable*. However, it's also possible that the mere existence of high-level variables is what influenced participants' memory for the presence of an explanation, or that the *potentially intelligible* condition would look more like the *intelligible* condition if it were certain that there were high-level variables to be discovered and that such high-level variables would indeed generate understanding.

The two sets of studies just described, Overton (2012) and Waskan et al. (2014), represent creative and important first steps in understanding scientists' and laypeoples' conceptions of explanation. The methods they develop can potentially be employed to address a variety of additional questions, such as the extent to which explanation is regarded as factive (see also Braverman et al. 2012), whether the criteria for what constitutes an explanation varies across contexts or domains, and how conceptions of explanation differ from those for related notions, such as description, prediction, empirical adequacy, and understanding.

3. Defining explanations functionally

The previous section considered folk and scientific conceptions of explanation and applications of the word "explanation." Another approach is to define explanations functionally, in terms of what they accomplish. For example, Lombrozo (2011) argues that although explanations can strike us as intrinsically valuable, they also play an important instrumental role in the discovery and confirmation of intuitive theories, which in turn support prediction and intervention (see also Lombrozo & Carey 2006). Along these lines, others have suggested that explanations should be understood in terms of their role in generating understanding (Achinstein 1983; Wilkenfeld 2014), supporting future judgments (Craik 1943; Heider 1958; Quine & Ullian 1970), or motivating the construction of causal theories (Gopnik 2000).

Research in psychology and education confirms that seeking, generating, and evaluating explanations can have important consequences for cognition (Lombrozo 2012). For example, explanations play a role in categorization (Ahn 1998; Lombrozo 2009; Murphy & Medin 1985; Rips 1989), in generalizing from known to novel cases (Rehder 2006; Sloman 1994; Lombrozo & Gwynne 2014), in resolving inconsistencies (Khemlani & Johnson-Laird 2013; Legare, Gelman, & Wellman 2010), in calibrating metacognition (Giffin, Wiley, & Thiede 2008; Rozenblit & Keil 2002), and in learning more generally (for reviews, see Lombrozo 2012; Fonseca & Chi 2010). But effects of explanation are not uniformly positive. For instance, generating explanations sometimes impedes learning about properties that are idiosyncratic (Williams, Lombrozo, & Rehder 2013) or irrelevant to the explanandum (Legare & Lombrozo, 2014; Walker, Lombrozo, Legare, & Gopnik, under review; Mishra & Brewer 2003). Documenting cases in which explanation has negative effects can be especially informative, not just because they have practical relevance (to pedagogy, for example), but also because such effects can be especially diagnostic of the mechanisms by which explanation influences reasoning. Just as visual illusions can reveal when and why visual perception is typically *accurate*, explanatory errors and "illusions" can help us identify when and why engaging in explanation is so often beneficial.

One challenge for a functional approach is to determine which consequences of explanation correspond to its proper function, and which are merely side effects. Most likely, explanations serve multiple functions, some resulting from the evolutionary origins of explanatory cognition, some from cultural evolution, some from learning, and some from the proximal goals of the agents who seek or employ them. To complicate things further, distinct functions are likely to compete: the best explanation for persuasion or efficient storage of information, for example, may not be the one that best supports future prediction.

Despite these challenges, a functional approach to explanation has two important advantages. First, if explanation serves one or more functions, one can ask what would *best* achieve those functions, providing a normative benchmark against which to assess psychological judgments and scientific practice. Providing such a benchmark also offers a potential bridge between descriptive, psychological claims about explanation and normative accounts of explanation from philosophy (Lombrozo 2011). Second, if explanatory judgments are at least moderately well matched to the functions of explanation, then the psychology of explanation provides important constraints on a functional account, and a functional account can in turn generate novel predictions about explanatory judgments and their cognitive consequences. Taking up this two-way relationship, the three sections that follow consider the basis for the perceived *quality* of an explanation, and what such judgments can tell us about what it is that explanatory judgments might be tracking. The first section considers formal measures of explanatory power; the subsequent two review empirical work on explanation in inference and learning.

4. Formal measures of explanatory power

Two recent papers have aimed to relate formal accounts of explanation to human judgment. Schupbach (2011) reports the results of an experiment in which participants were asked to evaluate the quality of two hypotheses (concerning which of two urns was selected) as explanations for different patterns of data (i.e., sets of balls drawn from one of the urns). Participants' responses were compared against five probabilistic measures of explanatory power – that is, of the strength of the explanation that a hypothesis provides for a specified explanandum. For instance, one measure was simply the extent to which a hypothesis h increases the probability of the explanandum d, i.e., $\mathrm{p}(d|h) - \mathrm{p}(d)$. Others were drawn from earlier work, such as Popper (1959), or from contemporary proposals by McGrew (2003) and Schupbach and Sprenger (2011). The results supported Schupbach and Sprenger's measure, $[\mathrm{p}(h|d) - \mathrm{p}(h|{\sim}d)] / [\mathrm{p}(h|d) + \mathrm{p}(h|{\sim}d)]$, as a better fit to human judgments than the alternatives he considered. Structurally, this measure corresponds to proposed measures of confirmation (Keynes 1921; Kemeny & Oppenheim 1952), which aim to capture the support that some evidence (=the explanandum) provides for some hypothesis (=the explanans).

Pacer et al. (2013) similarly compared formal measures of an explanation's quality against human judgments, but instead considered measures drawn from machine learning. These measures assessed variable settings for causes in a causal Bayes net as explanations for some observed effects. Pacer et al. presented participants with causal systems involving two or four causes and had them either generate or evaluate explanations for an observed effect. For both explanation generation and evaluation, they found moderate support for two models: the Most Relevant Explanation model (Yuan & Lu 2007), according to which the best explanation is the set of variable settings that has the highest likelihood, given the explanandum, relative to the summed likelihood of all other hypotheses, and the Causal Explanatory Tree model (Nielson et al. 2008), which – roughly – evaluates candidate explanations based on the log ratio of the probability of observing the explanandum given the candidate explanans to the summed probabilities of observing the explanandum under each hypothesis.

Notably, Pacer et al. (2013) found that a model that simply favors the most probable variable settings in light of the observations, Most Probable Explanation (Pearl 1988), faired quite poorly, in part because judgments were relatively insensitive to the prior probabilities of candidate explanations. Similarly, Schupbach (2011) found that his own measure provided a better fit to human judgments than the posterior probability of the corresponding hypotheses, which is itself a function of the prior. The findings thus suggest that human explanatory judgments track something more like evidence, information, or relevance, and not simply the prior or posterior probability of the explanans. This datum offers potential insight into the function(s) of explanation and how it might differ from processes that simply track probability. Indeed, as discussed in the sections that follow, explanation might play a special role in inference and in learning.

5. Inference to the best explanation

One candidate role for explanatory considerations is as a guide to inference. In particular, both scientists and everyday reasoners often appeal to the quality of an explanation as evidence for its value or truth, and explanatory considerations may well guide inference beyond the cases in which they are explicitly invoked. For example, explanations that are simpler may be remembered more effectively, related to evidence more easily, or simply strike us as more plausible than complex alternatives. Philosophers have debated the normative status of inference to the best explanation (Douven 2011; Lipton 2004), but systematic empirical work on the role of explanatory considerations in inference is relatively recent.

As characterized by Lipton (2004), inference to the best explanation (IBE) is an inference to the *loveliest* explanation. An important step, then, is recognizing what it is that makes an explanation lovely – that is, identifying and characterizing "explanatory virtues," or qualities that are thought to be desirable, such as an explanation's simplicity, breadth, or fruitfulness. As philosophers well know, however, each of these virtues is quite difficult to define, and it's likely that different virtues have a different normative status with respect to their role in inference. Here I provide a brief review of empirical work on simplicity and breadth, including their role in assessments of probability.

While appeals to simplicity in psychology are common across a wide spectrum of domains, from perception to language and development (Chater & Vitanyi 2003), the systematic investigations of simplicity in explanations likely began with Read and Marcus-Newhall (1993), motivated by Thagard (1989), and with Lagnado (1994). These studies quantified an explanation's simplicity in terms of the number of propositions or causes invoked, with simpler explanations involving a smaller number. Read and Marcus-Newhall found that participants preferred to explain sets of symptoms by appeal to one cause that explained all symptoms rather than two causes that jointly explained the same set. However, their studies did not control for effects of probability, leaving open the possibility that what appeared to be a preference for simplicity was actually a consequence of a preference for more probable explanations, given some reasonable assumptions about the baserates of different causes. Lagnado did take probability into account, and found that participants preferred a complex explanation over a simple one when the former was stipulated to be more likely. This suggests that if simplicity does inform explanatory preferences, it is trumped or made moot by probability. Finally, Lombrozo (2007) considered cases in which participants had to choose between one- and two-cause explanations, but where probabilistic evidence was communicated indirectly, by presenting the baserates for each cause, rather than explicitly telling participants which explanation was more likely. In this case, explanatory preferences were a function of simplicity *and* probability, with the basic pattern of results replicating with preschool-aged children (Bonawitz & Lombrozo 2012).

In more recent work, Pacer and Lombrozo (2014) have challenged the idea that a preference for one-cause explanations over two-cause explanations reflects a measure of simplicity that amounts to counting causes. In their initial experiments, they contrasted two different ways in which an explanation's simplicity could be evaluated: in terms of the total number of causes ("count simplicity"), or in terms of the total number of *unexplained* or "root" causes ("root simplicity"). To illustrate the difference between these measures, consider a patient who complains of both tiredness and weight loss. One could explain the tiredness by appeal to insomnia and independently explain the weight loss by appeal to a loss in appetite, thereby invoking two causes to explain the effects. Or it could be that depression is causing *both* insomnia and loss of appetite, which are in turn causing tiredness and weight loss. The first explanation invokes fewer causes total (2: insomnia + loss of appetite, versus 3: depression + insomnia + loss in appetite), and is therefore simpler according to count simplicity. But the second involves fewer "root" or unexplained causes (1: depression, versus 2: insomnia and loss of appetite), and is therefore simpler according to root simplicity. Using cases such as these, which pit count simplicity against root simplicity, Pacer and Lombrozo found that participants' preferences tracked the latter, not the former: the proportion of participants choosing a given explanation was a function of both the explanation's relative *root* simplicity and its relative probability (based on frequency information provided).

These findings by Lombrozo and colleagues suggest that simplicity plays a powerful role in explanatory judgments, even when independent bases for evaluation – such as frequency information – are available to the reasoner. In particular, simplicity affects an explanation's perceived "loveliness." But is loveliness also used to guide to "likeliness"? There's reason to think it is. In Bonawitz and Lombrozo (2012), children were asked what caused a set of observed effects, so participants' responses should have been a reflection of what they thought was most likely to be true, not (only) which explanation seemed most lovely. In Lombrozo (2007) and Pacer and Lombrozo (2014), however, participants were asked to identify the most *satisfying* explanations, not to make explicit assessments of which explanation was most likely. Nonetheless, both sets of studies also included an assessment of participants' *memory* for the frequency data that had been presented. These studies found that participants systematically misremembered the data in a way that rendered their explanation choice more likely than it actually was, but only under particular conditions: when the chosen explanation was simpler, and when the frequency information alone was insufficient to merit its selection. As a whole, the findings to date therefore suggest that in the face of even mild probabilistic uncertainty, people treat an explanation's (root) simplicity as one cue to its probability.

A few studies have also investigated effects of an explanation's breadth or "scope," where scope is understood as the number of observations or data points that the explanans can explain (for formal accounts of breadth or unification, see Myrvold 2003; Schupbach 2005). For example, Read and Marcus-Newhall (1993) found that people preferred explanations that accounted for a larger proportion of observed symptoms, but once again the study did not control for effects of (inferred) probability (see also Thagard 1989). Providing more indirect support for a preference for broad scope, Preston and Epley (2005) found that ideas that were used to explain a larger number of phenomena were judged more valuable to those who endorsed them, and Williams and Lombrozo (2010, 2013) found that prompts to explain can facilitate the discovery of broader patterns, as detailed in the section that follows. These findings point to the idea that explanations with broader scope are better, with an important qualification: when a causal explanation predicts effects that have an unknown status (i.e., it's not known whether or not they occurred), this "latent" scope seems to be considered a liability, not a virtue (Khemlani, Sussman & Oppenheimer 2011).

Clearly, many questions remain concerning the psychology of explanatory virtues and their role in inference. The findings to date, however, confirm the suspicion that explanatory considerations play a nontrivial role in everyday inference and reasoning. As this work moves forward, it will be valuable to relate it both to formal and to normative accounts of explanation from philosophy of science, formal epistemology, and other relevant disciplines.

6. Explanation and discovery

Explanation is closely connected with learning. We not only learn by receiving information in the form of explanations, but also through the very process of seeking and generating explanations (Lombrozo 2006, 2012). Understanding whether and how explanations influence learning is not only of practical relevance, but also an important constraint on both functional and formal accounts of explanation. Here I review a growing body of work relating philosophical accounts of explanation to the effects of explanation on learning (for a review of other accounts of explanation and learning, readers are directed to Fonseca and Chi 2010).

Many studies have compared the consequences of learning new material by explaining to oneself (without feedback) to those of engaging in a control task, such as thinking aloud or studying for a matched amount of time (Fonseca and Chi 2010). One consequence of "learning by explaining" could be that through explaining, the learner (implicitly) recruits a set of evaluative criteria for what constitutes a good explanation, and these criteria in turn influence how and what one learns (Lombrozo 2012). In particular, explanations may need to satisfy some formal constraints (e.g., being non-circular, or subsuming the explanandum under a more general pattern), and may be better to the extent that they exhibit explanatory virtues (such as low root simplicity and broad scope). These ideas have been explored most directly in the context of the "subsumptive constraints" account (Williams & Lombrozo, 2010, 2013), which focuses on subsumption and scope. Studies to date, involving both adults and preschool-aged children, find that a prompt to explain encourages learners to identify and generalize patterns with broad scope – i.e., those that apply to many cases.

In Williams and Lombrozo (2010), participants were tasked with learning to classify two kinds of novel robots from a set of eight exemplars. The exemplars were designed such that there was a subtle classification rule that accounted for all eight items, and also a more salient but imperfect rule that accounted for the category membership of only six robots (i.e., the rule had two exceptions). Participants who were prompted to explain why particular robots might belong to particular categories – without receiving feedback on their explanations – were significantly more likely to discover the perfect categorization rule than participants who described the robots, thought aloud while studying them, or engaged in free study. Subsequent work has replicated the effect with different explanation prompts (Edwards, Williams, Lombrozo, & Gentner 2014) and with preschool-aged children learning a novel causal relationship (Walker, Williams, Lombrozo, & Gopnik 2014).

If explanations are better to the extent that they apply to more cases, then those explanations that account for current *and past* observations should trump those that only account for current observations. On the assumption that one's current beliefs capture at least some prior observations (or are believed to have broad scope), this predicts that explanation could magnify effects of prior beliefs. And this appears to be the case: Williams and Lombrozo (2013) found that relative to participants in control conditions, prompts to explain led learners to preferentially discover and generalize patterns that were consistent with prior beliefs, and Walker et al. (2014) found similar effects with preschool-aged children learning about causal relationships.

In sum, the very process of explaining – even in the absence of feedback – can influence learning and inference, and at least some effects of explanation on cognition stem from the very properties of explanations that philosophers have identified, including a special role for broad, unifying patterns. One important question for future research concerns the relationship between these effects of explanation on individuals' learning and the role of explanations in science. For instance, how is theory construction and confirmation influenced by scientists' search for explanations, and when and why is this influence beneficial for scientific progress?

7. Explanation and the Negative Program in Experimental Philosophy

Thus far, we've been considering examples of what is sometimes called the "positive program" in experimental philosophy (see, for example, Alexander, Mallon, & Weinberg 2010), whereby empirical methods and results inform traditionally-philosophical questions, in this case concerning explanation. Another facet of experimental philosophy is the "negative program." In work exemplifying the negative program, empirical results are used to challenge the practice of relying on philosophical intuitions as evidence for or against particular philosophical claims. For example, some experimental philosophers have argued that intuitions about knowledge (e.g., Alexander & Weinberg 2007) and reference (e.g., Machery, Mallon, Nichols, & Stich 2004) are influenced by factors – such as the order in which vignettes are presented or one's culture, respectively – that should be irrelevant to the truth of the positions they're taken to support or refute. Such findings present a prima facie challenge to the reliability of philosophical intuitions as a guide to the concepts or claims they aim to elucidate (though they are by no means uncontroversial!).

For those engaged in the negative program, psychological work on explanation provides potentially lethal ammunition. That's because intuitions about the quality of explanations play a role well beyond research that focuses narrowly on explanation itself.

Indeed, many philosophical projects proceed by "inference to the best explanation," whereby the proponent for some position argues for its merits on the grounds that it provides the best explanation for some set of facts or intuitions (e.g., Paul 2012; but see van Inwagen 2009). If this argumentative step is itself suspect, then so are many philosophical conclusions.

There's certainly reason to believe that many of the factors that influence explanatory judgments are imperfect guides to truth. For example, people are more likely to side with the prosecution's "explanation" for a criminal act when the evidence is presented in an order corresponding to how events unfolded (Pennington & Hastie 1992). Novices are also less able to differentiate circular from non-circular explanations when irrelevant neuroscience is added (Weisberg et al. 2008). More generally, Trout (2002, 2007, 2008) argues that the "sense" of understanding provided by an explanation is often the result of psychological processes, such as hindsight bias and fluency effects, that are poor guides to whether the explanation furnishes *actual* understanding. If we agree that factors such as the order in which evidence is presented or the addition of irrelevant science should be orthogonal to the merits of an explanation, and if we have reason to think that they nonetheless influence judgments in actual philosophical practice, then empirical research on explanation might call into question a large body of philosophical work.

My own take on these issues is less pessimistic. First, the role of many such "errors" in explanatory reasoning could be negligible in philosophical practice. For example, the influence of evidence presentation order – which is thought to modulate the ease with which an explanation is constructed – is presumably reduced with time and reflection, although this prediction has not (to my knowledge) been tested directly. Other errors are eliminated with expertise: neuroscience experts, for example, don't fall prey to the "seductive allure" of irrelevant neuroscience (Weisberg et al. 2006; see also Erikkson 2012). More generally, whether and when explanatory considerations should factor into arguments is itself a question that has been subjected to philosophical scrutiny (e.g., Harman 1965; Lipton 2004). The greatest threat to the philosophical practice of IBE is therefore likely to come from "explanatory errors" that (a) are not moderated by reflection and expertise, and that (b) typically go unrecognized. It's not clear how often these conditions will obtain in the context of IBE (although there are almost certainly more subtle ways in which explanatory considerations pervade cognition, not always for the better). A final reason for optimism is that psychological research can help identify these subtle influences, potentially serving as a corrective to philosophical practice.

8. Towards an experimental philosophy of explanation

Despite rich theoretical work on explanation from philosophy, and diverse empirical work on explanation from psychology, the project of bridging these two fields is in its early stages. Notably, there have been few attempts to assess the correspondence between philosophical theories of explanation (i.e., causal, unificationist, pragmatic, etc.) and the explanatory intuitions of everyday reasoners as a way to evaluate the merits of such theories as descriptively adequate accounts of human psychology (for possible exceptions, see Chin-Parker & Bradner 2010 on pragmatic theories, Lombrozo & Carey 2006 on causal theories of teleological explanations, and Lombrozo, 2011, 2012 for more discussion). On the flipside, few philosophers have considered the empirical facts of explanatory cognition as a serious constraint on their theorizing, except insofar as those facts influence their own cognizing from the armchair.

Bridging these traditionally philosophical and empirical endeavors is itself a challenge in need of philosophical treatment. It's not clear whether the explanatory intuitions of preschool-aged children, for example, *should* constrain philosophical accounts. For starters, it depends on whether the accounts are regarded as normative or descriptive, and if the latter, on what their target is intended to be: explanations in everyday cognition, explanations in philosophy, explanations by scientists, explanations in the scientific corpus, etc. And to the extent that such accounts are regarded as normative, it's not clear what grounds their normativity. As suggested above, one attraction of a functional approach to explanation is that it could support an instrumental normativity: if we can identify the functions of explanation (for learning or for scientific discovery), we can then develop ideal models of what would best satisfy those functions (Lombrozo 2011). Formal approaches to explanation are likely to play an especially important role in this endeavor.

In sum, given recent developments in the cognitive and developmental psychology of explanation, in formal epistemology, and in both philosophical and psychological accounts of causation and understanding, the time seems ripe for an experimental philosophy of explanation.

References

Achinstein, Peter. 1983. *The Nature of Explanation.* New York: Oxford University Press.

Ahn, Woo-kyoung. 1998. "Why are different features central for natural kinds and artifacts?: The role of causal status in determining feature centrality." *Cognition* 69: 135-178.

Alexander, Joshua, Ronald Mallon, and Jonathan M. Weinberg. 2010. "Accentuate the Negative." *Review of Philosophy and Psychology,* 1: 297-314. DOI:10.1007/s13164-009-0015-2

Alexander, Joshua, and Jonathan M. Weinberg. 2007. "Analytic Epistemology and Experimental Philosophy." *Philosophy Compass,* 2: 58-60. DOI:10.1111/j.1747-9991.2006.00048.x

Bonawitz, Elizabeth B., and Tania Lombrozo. 2012. "Occam's Rattle: Children's Use of Simplicity and Probability to Constrain Inference. *Developmental Psychology,* 48: 1156-1164. DOI:10.1037/a0026471

Boumans, M. 2009. Understanding in economics: Gray-box models. In H. W. De Regt, S. Leonelli, & K. Eigner (Eds.), *Scientific understanding: Philosophical perspectives* (pp. 210–229). Pittsburgh, PA: University of Pittsburgh Press.

Braverman, M., Clevenger, J., Harmon, I., Higgins, A., Horne, Z., Spino, J., & Waskan, J. 2012. "Intelligibility is Necessary for Explanation but Accuracy May Not Be." *Proceedings of the Thirty-Fourth Annual Conference of the Cognitive Science Society*. Sapporo, Japan.

Bromberger, Sylvain. 1966. "Why Questions." In *Mind and Cosmos: Essays in Contemporary Science and Philosophy*, edited by Robert Colodny, 86-111. Pittsburgh: University of Pittsburgh Press.

Chater, Nick, and Paul Vitányi. 2003. "Simplicity: A unifying principle in cognitive science?" *Trends in cognitive sciences* 7: 19-22.

Chin-Parker, Seth, and Alexandra Bradner. 2010. "Background shifts affect explanatory style: how a pragmatic theory of explanation accounts for background effects in the generation of explanations." *Cognitive processing* 11: 227-249.

Craik, K. 1943. *The Nature of Explanations*. Cambridge, UK: Cambridge University Press.

Craver, C.F. and Lindley Darden. 2013. *In Search of Mechanisms: Discoveries Across the Life Sciences*. University of Chicago Press.

Douven, Igor. 2011. “Abduction.” In *The Stanford Encyclopedia of Philosophy (Spring 2011 Edition)*, edited by Edward N. Zalta. Accessed 05/01/2014. http://plato.stanford.edu/archives/spr2011/entries/abduction/

Dray, W. 1964. *Laws and Explanation in History.* New York: Oxford.

Edwards, B.J., Williams, J.J., Lombrozo, T, & Gentner, D. Under review. “Explanation recruits comparison: insights from a category-learning task.”

Eriksson, Kimmio. 2012. “The Nonsense Math Effect.” *Judgment and Decision Making,* 7: 746-749.

Fonseca, B. A. and M. T. H. Chi. 2010. “Instruction Based on Self-Explanation.”In *The Handbook of Research on Learning and Instruction*, edited by R. Mayer and P. Alexander (pp. 296-321). New York, NY: Routledge Press.

Friedman, M. 1974. “Explanation and Understanding.” *The Journal of Philosophy* 71: 5-19.

Griffin, Thomas D., Jennifer Wiley, and Keith W. Thiede. 2008. “Individual Differences, Rereading, and Self-Explanation: Concurrent Processing and Cue Validity as Constraints on Metacomprehension Accuracy.” *Memory & Cognition* 36: 93-103. DOI:10.3758/MC.36.1.93

Harman, G. 1965. “The Inference to the Best Explanation.” *Philosophical Review* 74: 88–95.

Heider, F. 1958. *The Psychology of Interpersonal Relations*. New York, NY: Wiley.

Hempel, Carl G. 1965. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press.

Hitchcock, Christopher. 2011. "Probabilistic Causation", *The Stanford Encyclopedia of Philosophy* (Winter 2012 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2012/entries/causation-probabilistic/>.

Kemeny, John G., and Paul Oppenheim. 1952. “Degree of Factual Support.” *Philosophy of Science* 19:307–24.

Keynes, John Maynard. 1921. *A Treatise on Probability*. London: Macmillan.

Khemlani, Sangeet, and P. N. Johnson-Laird. 2013. "Cognitive changes from explanations." *Journal of Cognitive Psychology* 25: 139-146.

Khemlani, Sangeet S., Abigail B. Sussman, and Daniel M. Oppenheimer. 2011. "Harry Potter and the sorcerer's scope: latent scope biases in explanatory reasoning." *Memory & Cognition* 39: 527-535.

Kvanvig, Jonathan. In press. “Understanding.” In *Oxford Handbook on the Epistemology of Theology* edited by Frederick D. Aquino and William J. Abraham. Oxford: Oxford University Press.

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

Legare, Cristine H., Susan A. Gelman, and Henry M. Wellman. 2010. "Inconsistency with prior knowledge triggers children’s causal explanatory reasoning." *Child Development* 81: 929-944.

Legare, Christine H., and Tania Lombrozo. 2014. “Selective Effects of Explanation on Learning in Early Childhood.” *Journal of Experimental Child Psychology*.

Lipton, Peter. 2004. *Inference to the Best Explanation, Second Edition*. London: Routledge.

Lombrozo, Tania. 2006. “The Structure and Function of Explanations.” *Trends in Cognitive Sciences,* 10: 464-470. DOI:10.1016/j.tics.2006.08.004

Lombrozo, Tania. 2007. “Simplicity and Probability in Causal Explanation.” *Cognitive Psychology,* 55: 232-257. DOI:10.1016/j.cogpsych.2006.09.006

Lombrozo, Tania. 2009. “Explanation and Categorization: How ‘Why?’ Informs ‘What?’” *Cognition,* 110: 248-253. DOI:10.1016/j.cognition.2008.10.007

Lombrozo, Tania. 2010. “Causal-Explanatory Pluralism: How Intentions, Functions, and Mechanisms Influence Causal Ascriptions.” *Cognitive Psychology,* 61: 303-332. DOI:0.1016/j.cogpsych.2010.05.002.

Lombrozo, Tania. 2011. “The Instrumental Value of Explanations.” *Philosophy Compass,* 6: 539-551. DOI:10.1111/j.1747-9991.2011.00413.x

Lombrozo, Tania. 2012. “Explanation and Abductive Inference.” In *Oxford Handbook of Thinking and Reasoning,* edited by Keith J. Holyoak and Robert G. Morrison, 260-276. Oxford: Oxford University Press.

Lombrozo, Tania, and Susan Carey. 2006. “Functional Explanation and the Function of Explanation.” *Cognition,* 99: 167-204. DOI:10.1016/j.cognition.2004.12.009

Lombrozo, Tania, and Nicholas Z. Gwynne. 2014. “Explanation and inference: Mechanistic and functional explanations guide property generalization.” *Frontiers in Human Neuroscience, 8*. doi:10.3389/fnhum.2014.00700

Machery, Edouard, Ron Mallon, Shaun Nichols, and Stephen P. Stich. 2004. “Semantics, Cross-Cultural Style.” *Cognition,* 92: B1-B12. DOI:10.1016/j.cognition.2003.10.003

McGrew, Timothy. 2003. “Confirmation, Heuristics, and Explanatory Reasoning.” *British Journal for the Philosophy of Science* 54: 553–67.

Mishra, Punyashloke, and William F. Brewer. 2003. “Theories as a Form of Mental Representation and Their Role in the Recall of Text Information.” *Contemporary Educational Psychology,* 28: 277-303. DOI:10.1016/S0361-476X(02)00040-1

Murphy, Gregory L., and Douglas L. Medin. 1985. "The role of theories in conceptual coherence." *Psychological Review* 92: 289.

Myrvold, Wayne C. 2003. "A Bayesian account of the virtue of unification." *Philosophy of Science* 70: 399-423.

Nielsen, Ulf, Jean-Philippe Pellet, and Andr´e Elisse- eff. 2008. “Explanation trees for causal Bayesian networks.”*Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Pacer, M., and Williams, J., Xi, C., Lombrozo, T., & Griffiths, T. L. 2013. Evaluating computational models of explanation using human judgments. *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI).*

Paul, L. A. 2012. “Metaphysics as Modeling: The Handmaiden’s Tale.” *Philosophical Studies,* 160: 1-29. DOI:10.1007/s11098-012-9906-7

Pearl, Judea. 1988. “Probabilistic reasoning in intelligent systems.” San Francisco: Morgan Kaufmann.

Pennington, Nancy, and Reid Hastie. 1992. “Explaining the Evidence: Tests of the Story Model for Juror Decision Making.” *Journal of Personality and Social Psychology,* 62: 189-206. DOI:10.1037/0022-3514.62.2.189

Popper, Karl R. 1959. *The Logic of Scientific Discovery*. London: Hutchinson.

Preston, Jesse, and Nicholas Epley. 2005. "Explanations Versus Applications The Explanatory Power of Valuable Beliefs." *Psychological Science* 10: 826-832.

Quine, W. V. O. and J. S. Ullian. 1970. *The Web of Belief*. New York: Random House.

Read, Stephen J., and Amy Marcus-Newhall. 1993. “Explanatory Coherence in Social Explanations: A Parallel Distributed Processing Account.” *Journal of Personality and Social Psychology,* 65: 429-447. DOI:10.1037/0022-3514.65.3.429

Rehder, B. 2006. “When Causality and Similarity Compete in Category-Based Property Induction.”*Memory & Cognition* 34: 3–16.

Rips, Lance J. 1989. "Similarity, typicality, and categorization." *Similarity and analogical reasoning*: 21-59.

Rozenblit, Leonid, and Frank Keil. 2002. “The Misunderstood Limits of Folk Science: An Illusion of Explanatory Depth.” *Cognitive Science,* 26: 521-562. DOI:10.1207/s15516709cog2605_1

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

Salmon, Wesley, C. 1989. *Four Decades of Scientific Explanation*. Minneapolis: University of Minnesota Press.

Schaffer, Jonathan. 2014. "The Metaphysics of Causation", *The Stanford Encyclopedia of Philosophy* (Summer 2014 Edition), Edward N. Zalta (ed.), forthcoming URL = <http://plato.stanford.edu/archives/sum2014/entries/causation-metaphysics/>.

Schank, R. P. 1986. *Explanation patterns: Understanding mechanically and creatively*. Psychology Press.

Schupbach, Jonah N. 2005. "On a Bayesian Analysis of the Virtue of Unification*." *Philosophy of Science* 72: 594-607.

Schupbach, Jonah N. 2011. "Comparing probabilistic measures of explanatory power." *Philosophy of Science* 78: 813-829.

Schupbach, Jonah N., and Jan Sprenger. 2011. “The Logic of Explanatory Power.” *Philosophy of Science* 78 (1): 105–27.

Sloman, S. A. 1994.“When Explanations Compete: The Role of Explanatory Coherence on Judgments of Likelihood.”*Cognition* 52: 1–21.

Thagard, P. 1989. “Explanatory coherence.” *Behavioral and Brain Sciences, 12*: 435–467.

Trout, J.D. 2002. “Scientific Explanation and the Sense of Understanding.” *Philosophy of Science,* 69: 212-233. DOI:10.1086/341050

Trout, J.D. 2007. “The Psychology of Scientific Explanation.” *Philosophy Compass*, 2: 564-591. DOI:10.1111/j.1747-9991.2007.00081.x

Trout, J.D. 2008. “Seduction Without Cause: Uncovering Explanatory Neurophilia.” *Trends in Cognitive Sciences,* 12: 281-282. DOI:10.1016/j.tics.2008.05.004

van Inwagen, Peter. 2009. “The New Anti-Metaphysicians.” *Proceedings and Addresses of the American Philosophical Association,* 83: 45-61.

van Frassen, Bas. C. 1980. *The Scientific Image*. Oxford: Oxford University Press.

Walker, C., Lombrozo, T., Gopnik, A, & Legare, C. Under review. “Explanation prompts children to favor inductively rich properties.”

Walker, C., Williams, J.J., Lombrozo, T., & Gopnik, A. In preparation. “The role of explanation in children’s causal learning.”

Weisberg, Deena S., Frank C. Keil, Joshua Goodstein, Elizabeth Rawson, and Jeremy R. Gray. 2008. “The Seductive Allure of Neuroscience Explanations.” *Journal of Cognitive Neuroscience,* 20: 470-477. DOI:10.1162/jocn.2008.20040

Weiskopf, D. A. 2011. Models and mechanisms in psychological explanation. *Synthese 183*: 313-338.

Wellman, Henry M. 2011. “Reinvigorating Explanations for the Study of Early Cognitive Development.” *Child Development Perspectives* 5: 33-38. DOI:10.1111/j.1750-8606.2010.00154.x

Williams, Joseph J., and Tania Lombrozo. 2010. “The Role of Explanation in Discovery and Generalization: Evidence from Category Learning.” *Cognitive Science,* 34: 776-806. DOI:10.1111/j.1551-6709.2010.01113.x

Williams, Joseph J., and Tania Lombrozo. 2013. “Explanation and Prior Knowledge Interact to Guide Learning.” *Cognitive Psychology,* 66: 55-84. DOI:10.1016/j.cogpsych.2012.09.002

Williams, Joseph J., Tania Lombrozo, and Bob Rehder. 2013. “The Hazards of Explanation: Overgeneralization in the Face of Exceptions. *Journal of Experimental Psychology: General*, 142: 1006-1014. DOI:10.1037/a0030996

Woodward, James. 2011. “Scientific Explanation.” *The Stanford Encyclopedia of Philosophy (Winter 2011 Edition)*, edited by Edward N. Zalta. Accessed 05/01/2014. http://plato.stanford.edu/archives/win2011/entries/scientific-explanation/

Woodward, James. 2013. "Causation and Manipulability." *The Stanford Encyclopedia of Philosophy* (Winter 2013 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2013/entries/causation-mani/>.

Yuan, Changhe and Tsai-Ching Lu. 2007. “Finding explanations in Bayesian networks.” In The 18th International Workshop on Principles of Diagnosis, pages 414–419.

Biographical Note

Tania Lombrozo is an Associate Professor of Psychology at the University of California, Berkeley, as well as an affiliate of the Department of Philosophy and a member of the Institute for Cognitive and Brain Sciences. Her research focuses on explanation, abductive inference, causal reasoning, learning, conceptual representation, and social cognition.