Sci & Educ
DOI 10.1007/s11191-015-9784-4

ARTICLE

Inference to the Best Explanation (IBE) Versus Explaining for the Best Inference (EBI)

Daniel A. Wilkenfeld[1] · **Tania Lombrozo**[1]

Abstract In pedagogical contexts and in everyday life, we often come to believe something because it would best explain the data. What is it about the explanatory endeavor that makes it essential to everyday learning and to scientific progress? There are at least two plausible answers. On one view, there is something special about having true *explanations*. This view is highly intuitive: it's clear why true explanations might improve one's epistemic position. However, there is another possibility—it could be that the *process* of seeking, generating, or evaluating explanations itself puts one in a better epistemic position, even when the outcome of the process is not a true explanation. In other words, it could be that *accurate explanations* are beneficial, or it could be that high-quality *explaining* is beneficial, where there is something about the *activity* of looking for an explanation that improves our epistemic standing. The main goal of this paper is to tease apart these two possibilities, both theoretically and empirically, which we align with "Inference to the Best Explanation" (IBE) and "Explaining for the Best Inference" (EBI), respectively. We also provide some initial support for EBI and identify promising directions for future research.

1 Introduction

In pedagogical contexts and in everyday life, we often come to believe something because it would best explain the data. We infer that the children made a snack because that best explains the mess in the kitchen, that a storm is coming because that best explains the rising winds, and that the organic avocados cost $1 because that best explains the sign

✉ Daniel A. Wilkenfeld
daniel.wilkenfeld@berkeley.edu

Tania Lombrozo
lombrozo@berkeley.edu

[1] Department of Psychology, 3210 Tolman Hall, University of California, Berkeley, Berkeley, CA 94720, USA

Published online: 08 October 2015

underneath them that says "$1." None of these inferences is certain, however, and each could be undermined by the introduction of a sufficiently good alternative explanation. If we know that there is a rogue prankster putting "$1" stickers on all organic produce, for example, then we are no longer justified in concluding that the organic avocados really cost $1. These inferences are governed by the general practice of Inference to the Best Explanation (IBE), which also features prominently in education: students infer to the best explanation when they come to accept a particular principle as the best account of a series of mathematical examples, or as they puzzle out the author's intention in writing a particular sonnet. Educators, too, infer to the best explanation when they infer that a student really does or does not understand a mathematical principle or a sonnet on the basis of how she responds to focused queries.

As endemic as IBE is in daily life, it is (perhaps) even more central to the advance of scientific knowledge. Among the entities of most interest to science are things that are too small, too far away, or too temporally distant to be observed directly (e.g., electrons or the Big Bang). Science also aims to characterize abstract entities or processes that cannot be observed because they have no specific location (e.g., economic inflation) and whose existence can only be inferred from the effect they have on observable objects (e.g., dark energy). There seems to be something valuable about this feature of science: the best scientific theories go beyond merely *describing* what we can directly observe to say something about what is behind a pattern of observations—what truly explains them.

These examples (and others like them) illustrate that explaining can foster deep knowledge of how the world works, but why? What is it about the explanatory endeavor that makes it essential to everyday learning and to scientific progress? There are at least two plausible answers. On one view, there is something special about having the right *explanations*. This view is highly intuitive: it's clear why true explanations[1] might improve one's epistemic standing—that is, one's accurate (or useful) beliefs about the world and one's potential for acquiring more. However, there is another possibility: it could be that the *process* of seeking, generating, or evaluating explanations can itself put one in a better epistemic position, regardless of whether or not the process results in an accurate explanation. In other words, it could be that engaging with *accurate explanations* improves one's epistemic standing, or it could be that *engaging in high-quality explaining* improves one's epistemic standing, where there is something about the *activity* of looking for an explanation that is itself beneficial. The main goal of this paper is to tease apart these two possibilities, both theoretically and empirically.

Of course, these two possibilities are not mutually exclusive. The process of explaining is sometimes beneficial because it leads to true explanations. We will argue, however, that there is another way explaining can potentially have epistemic value, arising from other consequences of engaging in explanation. We refer to this proposal as Explaining for the Best Inference (EBI),[2] and we contrast it with the better-known process of Inference to the Best Explanation (IBE). Insofar as IBE concerns itself with the process of explaining, it is only instrumentally so, as a means to acquiring true explanations. EBI, in contrast, allows for benefits that extend beyond that acquisition of a true explanation, and that can even coexist with the generation of a false one.

[1] For purposes of this paper, we treat "accurate," "true," and "right" explanations as roughly interchangeable. This begs certain questions—particularly as pertains to the debate about scientific realism and anti-realism—but these issues are not directly relevant to the distinction between IBE and EBI.

[2] This use of "EBI" is distinct from the use expounded in Persson (2007).

This distinction between the primacy of explanations and the primacy of *engaging in explaining* roughly maps onto the philosophical question of whether the primary locus of analysis regarding explanation is explanation (the product) or explaining acts. If the primary notion is explanation (the product), we would expect the value of explaining to be critically tied to the value of the product ultimately produced. If the primary notion is explaining (the process), this need not be the case, and a range of more subtle and indirect epistemic benefits might be considered. Note, however, that one could come to possess knowledge of a true explanation (the product) from any number of sources (e.g., reading it in a reliable book), not only from engaging in IBE.[3] In articulating and defending EBI, we will therefore adopt a relatively broad notion of IBE, as it is not critical—for our purposes—to differentiate between cases in which a true explanation is obtained via IBE (as described in the next section) and cases where it is obtained via some other process. If anything, taking a broader notion of IBE makes our defense of EBI stronger, as it must be differentiated from a broader range of cases.[4]

The structure of the paper is as follows. In the next section, we will go into more detail describing schematically what learning through explanation amounts to. We will present two versions of how explaining can lead us to make inferences; one model focuses on the inferential utility of learning true explanations (IBE), and the other focuses on the inferential utility of the act of explaining (EBI). Then, in Sect. 3, we consider how the processes of explaining and inferring are connected in two broad families of philosophical views of explanation: product-first accounts, which posit that something is an explanation if and only if it exhibits some particular structure or content, and explaining-first accounts, which identify explanations by the role they play in human discourse. Then, in Sect. 4, we turn to research from psychology that can help flesh out our distinction and provide support for EBI, and which seems to suggest that *both* explaining and hitting on correct explanations improve performance on epistemic tasks. In Sect. 5, we discuss why even explaining activities that do not result in true explanations might be epistemically valuable, with an examination of some proposed mechanisms by which the act of explaining could directly improve our ability to make inferences. In Sect. 6, we take a brief look at the question of whether the partial shift from explanation to explaining has any impact on the question of whether IBE or EBI are justified epistemic practices. We conclude with a discussion of the potential educational implications of recognizing EBI, including an examination of the connection between EBI and research on learning progressions.

Before we proceed, it's worth clarifying what it is we are and are not aiming to accomplish in this paper. First, although we will consider some mechanisms by which explaining affects learning and inference, it is not the purpose of this paper to articulate a process-level account of explanatory reasoning, in general or in EBI. Rather, our aim is to highlight and explore an aspect of explaining that is often overshadowed by its role in leading thinkers to true explanations. Second, our aim is not to provide criteria for identifying which acts of explanation "count" as EBI. On our view, any explanatory reasoning is a candidate instance of EBI. However, this does not entail that all instances of EBI are beneficial. People can be better or worse at explaining, just as they can be better or worse at

[3] This point was made to us by an anonymous reviewer.

[4] Even on our expanded notion, EBI and IBE are not quite exhaustive of the way explanations could be of cognitive value—it is possible that the product of explanations could be valuable for some reason aside from their truth, while also not as a function of the explanatory reasoning process that led to them. For instance, one could believe that having *any* explanation, regardless of its truth, is what's important. This view combines the focus on *explanations* from IBE with the tolerance for inaccuracy from EBI.

means-end reasoning. Our contention is that one can be good at explaining—and reap various cognitive benefits—even if one does not arrive at the correct explanation. As analogies, one might possibly think that reasoning could have benefits even if it does not result in valid arguments, and one might certainly think that painting could have benefits even if it does not result in good paintings.

2 IBE and EBI

2.1 Traditional IBE (Schematically)

We begin with a look at the nature of traditional IBE, in broadly schematic form. In the next section, we will plug in a family of philosophical accounts to see how it fills in the schema to provide a more precise picture of what IBE comprises.

"Inference to the best explanation" became a staple in the philosophical lexicon as a result of a 1965 article of the same name by Gilbert Harman. Harman pointed out that when we conclude that a set of facts admits of a best explanation, we naturally come to the further conclusion that that explanation accurately represents the world. To take his example, when the detective determines that the butler's having committed the murder best explains all the evidence, he will naturally conclude that the butler is the culprit.

There are several crucial aspects of IBE for examining its nature and justification. First, note that IBE is not (or at least not typically) a process of inferring to the *only* explanation. If proposition or phenomenon p really admitted of exactly one possible explanation, it would be relatively straightforward why we would infer that explanation to be true. After all, *something* (probably) explains p, so if the only candidate is E, then E is probably true. However—detective fiction notwithstanding—it is never (or almost never) the case that the explanation whose truth is being inferred is the only possible explanation. As has often been pointed out (perhaps most famously in Quine 1951), any particular proposition—including the proposition that some alternative explanation E_A explains p—can be maintained so long as one is willing to make (potentially drastic) corresponding changes to one's other beliefs. Thus, inference to the best explanation cannot ever show that an explanation *has* to be true, but (at best) that it is most likely to be true.

Peter Lipton, in his landmark (2004) exploration of IBE, observed two ways in which IBE could potentially be rendered trivial. First, on most accounts of explanation, it is assumed that E can only *really* explain p if E is true (or, if E is not a statement but a feature of the world, that a statement of E is true). Thus, as soon as one says that E is an explanation at all, one has committed to E being true, and so inferring E becomes trivial. Lipton's (natural) solution is to consider statements that meet all the criteria for being explanations except perhaps for truth, and then to assess their truth on the basis of their other features. Lipton calls these statements "potential explanations" and argues that IBE involves inferring that the best *potential* explanation is the actual (true) explanation.

Lipton's second and related point is that to avoid another sort of triviality, the likelihood of an explanation must not be considered as a virtue when we are engaged in the process of selecting the best explanation. To the extent we considered likelihood on a par with other explanatory virtues (such as precision and scope), IBE would be reduced (to that same

extent) to the process of inferring the *likeliest* explanation. Moreover, if we are allowed to consider the likelihood of an explanation, there seems little reason not to be good Bayesian thinkers and simply draw an inference in line with the probabilities. According to Lipton, we must instead consider an explanation's "loveliness"—the extent to which it exhibits all the virtues we look for in an explanation *except* antecedent likeliness—and conclude that the loveliest explanation is also the likeliest to be true.

Putting the pieces together, we see that IBE is the process of determining that one potential explanation out of many is the loveliest, and inferring from that that it is also the likeliest. So far, this framework remains neutral regarding what form explanations take, and hence what sorts of virtues they are likely to exhibit—that is, what it is that makes them more or less "lovely." In Sect. 3, we plug in a dominant family of accounts of explanation from the philosophy of science, and see what sort of picture it implies regarding the nature of IBE. But first, we highlight the ways in which IBE privileges *explanations* over the act of *explaining*.

2.2 IBE Versus EBI: Putting Explaining First

For our purposes, the most important feature of Lipton's view is that the primary locus of analysis is the explanation rather than the explaining act. It is the *explanation's* loveliness that is judged, and the *explanation's* likeliness that is determined. To the extent that the process of *explaining* is valuable, it is only instrumentally so, for how it brings us to the loveliest and eventually likeliest explanation. This is, we argue, a natural consequence of the bulk of modern philosophy of explanation. Most accounts are what we would classify as product-first—explanations are defined primarily, and explaining is only defined derivatively. If explanations are *conceptually* prior to explaining, then it is reasonable (though not inevitable) to suppose that they are *evaluatively* prior as well—that is, that any value of explaining derives from the relevance and truth of the particular explanation inferred. By contrast, there is a minority position in the philosophy of explanation—which finds some sympathy in the psychology of explanation—that treats explaining *acts* as the unit of analysis. This makes room for the possibility that explaining can be valuable independently of the explanations it happens to produce, and in particular how accurate those explanations are.

An analogy is perhaps helpful. Most traditional accounts of meaning took sentences (or sometimes words) to be the primary unit of meaning. On such views, speaking is only meaningful—and thus, insofar as our concern is with meaning, only valuable—to the extent that it leads to sentence production and consumption. By contrast, in the twentieth century philosophers began to focus their attention on the productive nature of language. The thought, first put forward in Wittgenstein's *Philosophical Investigations* (2001/1953), is that language use is something people *do*, and hence sentences are best understood as byproducts of communication and other uses of language. This led to a massive reorientation in the philosophy of language regarding what made languages communicatively valuable. Suddenly, language *use* could be valuable in ways independent of the content of sentences. Similarly, exploring views that focus on *explaining* rather than just explanations might suggest a reorientation in the philosophy of science regarding what makes explanations epistemically valuable—that is, how and what they enable us to learn about the world. Such a reorientation might also have practical implications, which we turn to in Sect. 7.

3 Good Explanation Versus Good Explaining in Philosophy

Within the philosophical literature, the vast majority of the work on explanation takes the product as the basic locus of analysis. In this section, we review one family of examples of traditional approaches to explanation in philosophy—causal accounts—and examine what assumptions they would have to make about how the process of IBE works and what its cognitive benefits are. We focus on causal accounts because they are arguably the most popular accounts in the philosophy of science today, and they seem to be the accounts most often assumed by psychologists. However, what we say about the connection between causal explanation and IBE will apply, with appropriate modifications, to deductive-nomological and unificationist accounts of explanation as well.[5]

3.1 Causal Accounts of Explanation

According to causal accounts, we explain a phenomenon by identifying its causes. Given philosophers' hesitation to treat causation as an unanalyzed notion (this dates back at least to Hume 2000/1748), much of the work of causal theorists has been directed at spelling out in more detail what causation amounts to. This generally takes the form of either a process account (e.g., Salmon 1984), a counterfactual dependence account (e.g., Woodward 2003), or an account in terms of regularly functioning hierarchically composed mechanisms[6] (e.g., Craver 2007).[7] For present purposes, the details of what causation comprises do not matter; the general picture—that to explain is to identify one or more causes (whatever that means)—is the same on all accounts. Intuitively, someone who tells us what *caused* the car crash has *explained why* the car crashed.

3.1.1 Causal Accounts and IBE

On a causal account of explanation, one infers to the best explanation by identifying the potential cause(s) of a given phenomenon that would provide the loveliest explanation. Lipton himself advocates a causal view, and one can see why it is a natural fit with inference to the best explanation. Since causal relations are (presumably) real features of the universe, identifying a phenomenon's cause straightforwardly tells you something about what the world is like. Moreover, causes are a particularly important part of the world to identify, as they are typically the loci of possible interventions (Woodward 2003).

Causal accounts of explanation find natural partners in the extensive literature on causal inference, which is also a boon for IBE. For example, Lipton cites as good resources for causal inference both Mill's methods for identifying causes (2004, p. 18) and more modern Bayesian confirmation theory (2004, Chapter 7). If one can show that features of loveliness (which for Lipton includes an array of virtues such as precision, scope, mechanism; 2004, p. 122) track one or both of these guideposts to causation, one has gone some way to not only describing IBE, but justifying it as well. More precisely, one has gone some way toward reducing the problem of justifying IBE to the more general Humean problem of

[5] On the former account, IBE helps us by pointing us to the laws and initial and boundary conditions governing a system, and on the latter account, it helps us by pointing to assumptions and argument patterns that would unify our overall knowledge store. (For a review of those accounts, see Woodward 2014).

[6] Mechanists rarely claim that mechanisms account for all causal relations, though.

[7] To be clear, this taxonomy of causal accounts is not meant to be exhaustive; as the details of the accounts are not relevant for our analysis, a comprehensive review is unnecessary.

how to justify *any* of our causal inferences, which, while still intractable, is a problem for anyone else who takes the scientific process seriously, not just for Lipton and other advocates of IBE.

Importantly, there is nothing about the actual *process* of explaining that is epistemologically valuable on the causal account. The cognitive upshot of engaging in causal IBE comes from identifying actual *explanations*, i.e., causes in the world. If one hits upon the wrong causes, one has failed in the explanatory endeavor, whatever processes one might have employed to reach one's conclusions. Conversely, if one has hit upon genuine causes, one has gained what cognitive benefit there is to be had from causal IBE. The relevant causes or statements of causes are static entities—they could be derived for oneself, but they could just as easily be found in books. Inferring that E is the explanation of p is valuable because the content of E says something about what the actual causal structure of our world must be. The process of finding the right explanation is entirely instrumental. In an empirical hypothesis with echoes of this view, Alison Gopnik (1998) suggests that we have evolved to value explanations precisely because of their instrumental role in motivating us to identify the causal structure of the world.[8]

3.2 Explaining-First Accounts of Explanation

In contrast to "explanation-first" accounts, there is a class of accounts of explanation that focuses on the act of explaining as the primary unit of analysis, with explanations defined derivatively. For instance, Peter Achinstein (1983) argues that "explainings" are a particular type of speech act, such that one explains by producing an utterance that attempts to cause understanding in some particular audience (1983, p. 52). Note that explaining for Achinstein is a fundamentally situated activity, grounded in real audiences' interests and background presuppositions. Indeed, the explanation itself falls out of Achinstein's theory entirely, to the point that he baptizes his view a "no-product view." He writes:

> Any sentence containing an explanation product-expression is paraphrasable into a sentence that contains no such expression, but does contain one or more expressions of the form 'a is an act in which S explained q by uttering u.' When this thesis is combined with the view that such paraphrases are more fundamental than sentences with product-expressions we get the no-product view (Achinstein 1983, p. 98).

Similarly, Wilkenfeld (2014) argues that explanations are best characterized by their connections to explaining acts, which are in turn best characterized by their propensity to produce understanding. A central tenet of Wilkenfeld's (2014) view is that "what makes something an *explanation* is its being content which, if true, could, under appropriate circumstances, be used in *explaining-acts*" (3370). While this view is grounded in an assumption that explanations foster genuine understanding—and so are in some sense objective—the explaining act is nonetheless conceptually prior to any actual explanations.

3.2.1 Explaining-First Accounts and IBE/EBI

The important point, on both Achinstein's and Wilkenfeld's accounts, is that one cannot make sense of the notion of explanation independent from actual, contextually situated

[8] It is worth pointing out that the remarks about the unimportance of the *act* of explaining extend to any theory that takes the explanation as the only valuable product of the explaining act. In other words, everything we have said about IBE in the narrow sense applies to the broader class of explanation-acquisition discussed in Sect. 1.

explainers and explanation-requests. Explanations are, on these views, what people use to accomplish explaining. While explaining is a peripheral and eliminable notion on other accounts (IBE must, as Lipton suggests, involve an inference based on the content of the explanation itself), these alternative accounts create space for a radically different form of learning or inference through explanation—one in which explainers and explaining acts are in the foreground. Perhaps one can learn not only from inferring the best *explanation* (as determined by content), but also by engaging in the best *explaining act*.[9]

One might suspect that, in most cases, the best explanation will go hand in hand with the best explaining act. They can, however, occasionally diverge. It is possible that there are virtues to an explaining act that are not themselves virtues of explanation. Since explaining acts are events in real time, we can look at their causal effects in a way that would not make sense for the abstracta that are more typical explanations. One particular effect of explaining might be that it directs the mind of the explainer to particularly useful features or dimensions, perhaps causing as much cognitive gain in the explainer as in the recipient of the explanation (we discuss other possible boons of explaining in Sect. 5). It might then be epistemically responsible for agents to allow themselves to be guided by what would feature best in an *act of explaining* (though see Sect. 6), even if that would not provide the best explanation per se. This view predicts that there could be epistemic value in explaining even if one ultimately advances an inaccurate explanation.

3.2.2 An Aside on Erotetic Accounts

It is worth differentiating accounts like Achinstein's and Wilkenfeld's from "erotetic" accounts, which share some superficial similarities. According to erotetic accounts, explanations are answers to why-questions, though accounts differ in what constitutes a (good) answer. On Sylvain Bromberger's (1966) account, the issue is entirely formal: one needs to explain when one has identified a deviation from an otherwise general law, and one *successfully* explains when one has located the basis of that deviation. For example, one explains why this tree did not shed its leaves by noting that, while most trees do, this tree is an evergreen. Garfinkel (1980) and van Fraassen (1980) both emphasize the role of the contrast class, arguing that we only explain when we have determined why an event happened as opposed to some alternatives stipulated (or implied) in a question. One does not explain facts in a vacuum, but only relative to particular interests and contrasts. At least in the case of van Fraassen, the account is often characterized as being pragmatic—which may connote the importance of explaining as a situated activity—so it is worthwhile to emphasize the structural differences between erotetic accounts and what we call explaining-first accounts.

Looking at erotetic accounts, one finds that they do not actually promote the virtues of the explaining *activity*. While accounts such as van Fraassen's (1980), Garfinkel's (1980), and Bromberger's (1966) all contend that explanations answer questions, it is the explanation's formal properties that are relevant to its quality, not the process of explaining. Put differently, the relation between question and answer is entirely third-personal; for any given question there is a correct set of answers, and the benefit one gets from inferring a

[9] Achinstein and Wilkenfeld's accounts are still open to the possibility that the best explanations are the best as a result of having some particular explanatory virtues (on Wilkenfeld's 2013, representation-centric approach, natural candidates would be the accuracy and fecundity of the representational content of the explanation). Thus, their accounts do not preclude evaluating explanations in terms of their more traditional virtues, but do allow for an additional class of considerations.

good explanation is that one has identified one of those right answers. Process-first accounts, by contrast, take the situated explaining act as primary.

4 Explaining in Psychology

Within psychology, there is good evidence that people engage in "explanation-based reasoning" that resembles IBE (Lombrozo 2006, 2012; Hastie and Pennington 2000). For example, Lombrozo (2012) argues that engaging in explanation recruits evaluative criteria for what constitutes a *good* explanation (e.g., simplicity, scope), and that these criteria in turn constrain learning and inference, in part by leading people to discover and generalize the "best" explanation (for evidence, see Bonawitz and Lombrozo 2012; Legare and Lombrozo 2014; Lombrozo 2007; Walker et al. 2014; Williams and Lombrozo 2010, 2013; Williams et al. 2013). Pennington and Hastie (1986, 1988, 1992) similarly find that the decisions of mock jurors are influenced by the quality of the defense's and the prosecution's "explanation" for the evidence, where the explanation's coverage and other virtues inform assessments of quality and hence probability. While broadly consistent with IBE, however, this work was not designed to differentiate IBE from EBI. Because IBE and EBI will so often go hand in hand, it's useful to consider a situation in which they will likely diverge: when engaging in "good" explaining generates "bad" (that is, false) explanations. If such cases can nonetheless yield epistemic benefits—such as better inferences—then we have prima facie evidence for EBI.

In fact, recent evidence suggests that, at least in children, attempts to explain can sometimes serve a valuable epistemic role even when the explanations the children hit upon are false. In particular, Walker et al. (2014) explore the role of prompts to explain in addressing the question: how do young children generalize properties from one object to another–on the basis of non-obvious causal affordances or on the basis of superficial similarities of appearance? In one study, children were presented with trios of blocks: one that looked a certain way and activated a toy, one that looked similar but did not activate the toy, and one that looked different but did activate the toy. Children were then shown a previously hidden internal component of the first block and asked whether it was more likely to be shared by the perceptually similar block or the causally similar block.

Between the ages of about 3 and 6, children undergo a developmental shift: younger children tend to generalize on the basis of perceptual similarity, whereas older children (and adults) tend to favor causal similarity (e.g., Sobel et al. 2007). The question Walker and colleagues asked was whether (and how) being prompted to explain why each block did or did not activate the toy would influence the nature of this inference. The study revealed that prompts to explain had a systematic effect on performance. Those children who were prompted to explain why each block did or did not activate the toy were significantly more likely than those in a control condition to generalize on the basis of causal similarity over perceptual similarity—that is, they exhibited the more mature (and arguably more correct) pattern of inference, even though they did not receive any feedback on the quality or accuracy of their explanations.

The proponent of traditional IBE (at least in the broader sense of explanation-acquisition described in Sect. 1) has a ready explanation of these data. There is some best explanation of why an object has a particular causal role—in this case, an explanation that appeals to internal components of the block. Children who are prompted to explain generate an explanation that appeals to internal components and are subsequently more likely

to believe that internal components explain the objects' effects. They then (presumably unreflectively) invoke the maxim that like effects have like causes, and so conclude that the block that produces the same effect is likely to share the same internal components.

For present purposes, the important step in this account of children's improved inferential capacities is their realization that the likely explanation of the first block's causal profile is its internal components. Without this step, the chain of inference that leads them to the correct generalization cannot even get started. This might seem a reasonable and even necessary part of the account of children's improved performance—how could one get to generalizing internal components from one block to a causally similar block except by way of figuring out that the first block's effects are a result of internal components? The remarkable thing is that even children who opted for some *other* explanation of the first block's effect (for instance, one that appealed to kind membership) were more likely to generalize internal components to the causally similar block than children who weren't prompted to explain. The authors summarize:

> We also found evidence that the prompt to explain impacted children's inferences even when the explanations that were generated did not appeal to internal properties. For example, the two children who provided no modal explanation (i.e., children who provided distinct explanation types for each set) and the two children who provided a modal explanation of "no guess" were (numerically) the most likely to select the causal match (88 % each). In fact, each category of modal explanation, regardless of type (appearance: 53 %, kind: 63 %, other: 45 %), was associated with a higher proportion of causal matches than that observed of children in the control condition (40 %). (Walker et al. 2014, p. 348)

Note that even the children who explained the first block's performance in terms of its *appearance* more often generalized internal components to a block that was similar in causal role (over one that was similar in appearance) than did children in the control condition. This result is baffling on the traditional picture of IBE without EBI. If children explain the block's causal properties by appeal to appearance, then there would be no basis for expecting them to generalize internal features on the basis of causal rather than perceptual similarity.

A proponent of EBI has an alternative prediction: while hitting on the right explanation is the most surefire way to make the right inferences, the very process of explaining could yield cognitive benefits for other reasons, and so guide us to the right answer indirectly. What might these cognitive benefits look like, and how might they arise? We now turn to a more speculative discussion of what the relevant mechanisms underlying the benefits of explaining might be.[10]

5 The Benefits of the Wrong Explanation

There is great intuitive pull to the idea that explaining leads one to make good inferences only to the extent that it also leads one to the correct explanation; therefore, the results discussed in Sect. 4 might seem paradoxical. In this section, we review several possible mechanisms that could account for why even imperfect attempts to explain (imperfect in that they lead to false explanations) could advance one's epistemic standing and improve

[10] Of course, hitting on false explanations won't *always* be beneficial; conditions that have a net epistemic benefit when explanations are false may be rare, even if they are important in providing evidence for EBI. For instance, we know that in some cases explaining can actually impair learning, and that this appears to be in part because people perseverate in making judgments on the basis of false explanations. See for example Williams et al. (2013) on how prompts to explain in uncooperative worlds can actually impede learning.

performance on subsequent tasks. Our aim in reviewing such mechanisms, and in pointing to relevant data, is merely to suggest the possibility that there exist some such processes that lead to cognitive benefit *independently of their leading people to a true explanation*. We do not mean to imply that these examples constitute all and only examples of EBI, that they are sufficient to articulate a process-level account of EBI, or that they are all governed by some common underlying process.

A first way in which explaining acts that ultimately appeal to incorrect explanations (henceforth "imperfect explainings") could be beneficial is by focusing attention on certain features of objects and events that are causally relevant, regardless of whether these features are correspondingly privileged in explicitly formed explanations. Broadly consistent with this idea, Legare and Lombrozo (2014) find that prompts to explain lead children to do better on subsequent measures of causal learning, but worse on measures of memory for causally irrelevant details. For instance, children who are prompted to explain how a gear system works are better at identifying which novel gear can replace a missing piece to make the system operational, but less accurate in identifying which novel gear matches a missing gear's color. This suggests that the act of explaining directs attention and other cognitive resources to some stimuli over others—in this case, perhaps, to the size and shape of gears rather than their color. Returning to the findings from Walker et al. (2014), the idea is that in the process of explaining a child foregrounds the important similarity of causal role in a way that leads her to correctly generalize internal structure on that basis, even when the ultimate causal understanding is absent or misguided. The general idea is that explaining could affect one's focus, even if it doesn't ultimately cause one to form the correct occurrent belief regarding the causal structure of the world.

A second way in which imperfect explainings could help is by promoting abstraction. Williams and Lombrozo (2010) found that adult participants who were prompted to *explain* the category membership of novel objects named more abstract features of those objects than did those who were prompted to *describe* the objects. Specifically, in these studies, the objects were novel robots that varied in foot shape and in color (among other features). Participants who were prompted to explain were more likely to describe the foot shapes as "pointy" or "flat," as opposed to naming particular shapes (triangle, circle, etc.), and to describe the colors as "warm" or "clashing," as opposed to naming particular colors (red, blue, etc.). Interestingly, this was found for both foot shapes and color, even though foot shape was the only feature relevant to the categorization task. This suggests that, in situations where more abstract ways of thinking better promote one's epistemic aims, one might be well served by being in an "explaining-mindset," even if it does not ultimately lead one to accurate explanations.

A third way in which imperfect explainings could be cognitively beneficial is by prompting comparisons between analogous cases. Consider a set of studies by Edwards, Williams, and Lombrozo (2013), in which adult participants were again tasked with learning to categorize novel robots. Some participants were prompted to *explain* the category membership of training examples, and others to *compare* training examples to each other. In addition, at the end of the task, participants were asked to report how much comparison they engaged in spontaneously (i.e., whether or not they were instructed to do so). The researchers found that prompts to explain not only increased the rate at which participants discovered a key categorization rule, but also increased the extent to which participants reported engaging in comparison. This suggests that one feature of explaining is that it prompts people to compare relevant cases; such comparisons could themselves foster abstraction (Gentner and Markman 1997) and could also lead to insight independently of arriving at the correct explanation of a given feature in a particular case.

A fourth way in which imperfect explainings could help is by bringing to light the conflict between the content of the explanation and other beliefs or evidence. Chi et al. (1994), for instance, found that students learning about the circulatory system from a text often generated inaccurate "self-explanations," but that these explanations were not detrimental to learning. In the study, "self-explanations" were defined as the product of "spontaneously generating explanations to oneself as one studies worked out examples from a text." (Chi et al. 1994 p. 440). They write:

> It is even conceivable that generating incorrect self-explanations can provide a learning experience. Here is one possible interpretation. Having articulated an incorrect explanation, a student continues to read the next sentence or sequence of sentences in the text. Eventually, the text sentences, because they always present correct information, may contradict knowledge embodied in the incorrect self-explanation. This will create a case of conflict. Hence, one interpretation is that creating an incorrect self-explanation merely objectifies that piece of knowledge, which allows it to be examined in the face of conflicting information from subsequent sentences, thus establishing the opportunity for self-repair to resolve the conflict. (Chi et al. 1994, p. 471)

In other words, explaining could make a commitment explicit, facilitating the (ultimately beneficial) rejection of that commitment in the face of more reliable information. Losing faith in faulty understanding might in some circumstances be just as valuable as gaining more true beliefs—even if one never fully arrives at the truth. If I lose my confidence that I understand how to play poker, that can save me a lot of money (in games not played) even if it never leads me to gain an accurate understanding of strategy in poker.

Relatedly, imperfect explaining could lead one toward the truth by increasing metacognitive awareness of when one has a failure of comprehension. McNamara (2004) found that training in self-explaining (in the form of watching a video of someone self-explaining and then self-explaining on practice trials) increased students' propensity to flag when they did not understand something (p. 23), which can be an important step in learning. In another study, Rozenblit and Keil (2002) examined the extent to which people accurately rate their own understanding of how something works, such as a toilet or a zipper. They found that people significantly overrated their own understanding, but that this metacognitive error was significantly tempered simply by being prompted to provide an explanation of how the thing worked (p. 9—the relevant element is the decline in self-assessed understanding between "T1" and "T2", which are separated only by an attempt to explain). This suggests another possible benefit of explaining: to bring to light the otherwise hidden "illusion of explanatory depth" that blinds people to the gaps in their own understanding (Rozenblit and Keil 2002). Again—losing bad views might be as helpful in some circumstances as gaining good ones.

Of course, there is another obvious cognitive benefit of hitting on the incorrect explanation, which is that it in some way might take one closer to the true explanation, perhaps by causing one to rule out other false explanations. For example, imperfect explainings could succeed in virtue of the fact that, in highlighting a bad explanation, they at some level bring to light its failings. (Seeing the failure of our own views might be thought of as another form of metacognitive awareness.) Explaining could lead us to reject a pernicious false view, which then gets us part of the way to the true one. Such rejection seems to be what is at work in the cases discussed in Chi et al. (1994) and Rozenblit and Keil (2002), as people face the inadequacy of their own explanations and are then exposed to more accurate information. In fact, one might harbor a suspicion that *all* of the possibilities outlined above are only derivatively useful, insofar as they bring us closer to the correct explanation.

But critically, this need not be the case. Each example we considered could lead to benefits even before an accurate explanation is acquired, and even if such an explanation is *never* acquired. One might reasonably contend that EBI can only be cognitively beneficial insofar as it brings us closer to *some* truth, but the relevant truth need not be the content of the explanation, and this is where EBI departs from IBE. Being brought closer to the truth about, for example, the appropriate generalizations to draw from a set of observations, is logically independent from coming closer to the truth about a particular explanation or the facts to which that explanation adverts (even if, in practice, all three are often intertwined).

Finally, engaging in the (imperfect) process of explaining could promote a more sophisticated understanding of the nature of science and the acquisition of knowledge more generally. To borrow terminology from Duschl (1990), presenting scientific explanations as best evaluated in terms of their ultimate correctness treats science as a "final form" whose value is in its pronouncements. This perspective, however, could lead people to treat the pronouncements of science as no different—and hence no better—than the pronouncements of an oracle: "Such models unduly emphasize the successes of science and the final form of scientific knowledge" (Duschl 1990, p. 82). Duschl also argues that it is for this reason that people think creationism should be given equal attention to evolution—if they are both taken as mere final pronouncements, why value one over the other? Relatedly, Lehrer and Schauble (2006, p. 159) note that teaching "theories as final form science may leave students in the dark about the way knowledge is generated and may also distort the nature of scientific knowledge, inappropriately conveying that it is unchangeable and uncontested." By contrast, when one is engaged in the activity of explaining, one might be acutely aware of the possibility for error and improvement; this is even more the case if one becomes aware of how explaining can be imperfect. Even and perhaps especially when one never arrives at a correct explanation, explaining situates one within the *process* of science.

The various mechanisms just considered are not mutually exclusive, nor are they exhaustive in identifying possible benefits of engaging in explanation. It is likely that the interaction of all the above (and more) bring about positive inferential utility from explaining acts, even when those acts result in false explanations. The broader idea is that the mere act of explaining affects which aspects of the environment we attend to (e.g., causally relevant inner parts), at what level of abstraction we represent those aspects (e.g., "triangular" vs. "pointy"), and how we engage with them (e.g., by comparing them to each other). It also affects how we use what we learn to prepare ourselves for new data, and how we monitor our own comprehension as we go.

The potential importance of these mechanisms prefigures a response to another concern about EBI—that by jettisoning the connection between explaining and the truth of the explanatory product, we have no basis for evaluation: "anything goes." However, the benefits discussed above suggest one way different types of EBI can be better or worse compared to others even if the criteria for evaluation are distinct from (but overlapping with) those of IBE. Instead of truth of the explanation, we assess something like the utility of its consequences. This is admittedly a more difficult type of success to measure, but research to date already suggests various good-making features of explaining activities. For instance, we have reason to think that some of the processes engaged in explaining—such as abstraction and comparison—are useful practices. One could develop a metric for explaining-quality based on the extent to which a given act of explaining successfully utilizes these processes. Moreover, some of the same desiderata for good explanations will also apply to explaining acts. All else being equal, explaining acts where the hypotheses

considered are more internally consistent and antecedently plausible (see the discussion in Sect. 7 of "internal coherence") will be more successful acts.

6 EBI, Normatively

This article has primarily been concerned with the question of how people use explanations to make new inferences and generalizations—that is, with how learning through explanation is best *described*. There is an equally compelling question regarding whether it is wise to use explanations in this way—that is, whether our inferential practices are *justified*. This has proven to be a philosophically tricky issue. The central problem for traditional IBE is what Lipton refers to as the "Voltaire objection" (Lipton 2004, p. 143)—why think the virtues that guide assessments of explanation quality also guide truth? For example, why think the world is simple, consilient, or functions on the basis of understandable causal mechanisms? One could try to appeal to the track record of past success that our previous inferences to the best explanation have had, but—aside from worries about the pessimistic induction that most of our best theories have turned out to be false—the inference from the success of a practice to a conclusion that it generally identifies truth-tracking features seems suspiciously like an inference to the best explanation of that success.[11] Problems with this structure are very familiar from the history of philosophy—the Voltaire objection could be seen as a generalized version of the famous Humean objection that questions how we can ever be justified in reaching conclusions that go beyond the already-observed evidence.[12]

While the Voltaire objection poses problems for traditional IBE, one might hold out hope that the shift from a focus on explanation to a focus on explaining acts would resolve the dilemma. After all, we need no longer assume that statements with particular virtues necessarily reflect the truth, and so the Voltaire objection, as originally stated, cannot arise. This is cold comfort, however, for it is easy to raise an analogous concern for EBI. One contention of EBI is that the very act of explaining focuses our attention on aspects of a problem that are the basis for sound inferences and generalizations. But what aspects are those? On the one hand, the proponent of EBI could refuse to specify at all what features of our reasoning-while-explaining get us any closer to the truth—but that would be unhelpfully mysterious. Alternatively, the proponent could argue (as in Sect. 5) that EBI causes us to track certain virtues, modes of representation, comparison to other cases, etc., and further argue that such a focusing effect moves us toward the truth. However, we would only be justified in believing that tracking features such as simplicity (a paradigmatic explanatory virtue) moves us toward truth if we are already justified in believing that simplicity correlates with truth—the impossibility of justifying which forms the crux of the Voltaire objection. Similarly, abstraction only moves us toward the truth if the world functions under appropriately abstract principles, comparisons only move us toward the truth if like problems have like solutions, and so on.

[11] One anonymous reviewer pointed out to us that several thinkers (e.g., Papineau 1993 Sect. 5.11) have argued that this sort of rule circularity is perfectly reasonable. If this is right, then there is no problem for EBI either, and so much the better. However, our own inclination is that such seemingly question-begging responses should be adopted only as a last resort (though perhaps "last resort" accurately reflects our present place in the dialectic).

[12] Briefly: such justification cannot advert solely to the *a priori*, since our conclusion has to do with the actual contingent arrangement of the world, but it cannot be based on prior experience, since the validity of extending lessons learned from prior experience to unseen cases is the very thing in question.

There are responses to these concerns. One could argue individually for the rationality of favoring each virtue, as seen—for instance—in Bayesian arguments for favoring simpler hypotheses (e.g., Jefferys and Berger 1992). Alternatively, one could argue that such concerns just erode higher-order justification—our justification in believing that our beliefs are justified—without adversely affecting our first-order justification in the conclusions of our explanation-based reasoning. Our general contention here is the more modest one that such solutions are relatively insensitive to the question of whether our thoughts are guided by IBE or EBI.

That being said, there are some advantages with regard to the normative problems that are gleaned from refocusing our attention from explanation to explaining. When explaining is seen as a real activity undertaken by thinkers in the world, it is evaluable along practical dimensions. This offers a different *kind* of evaluation, independent of an explanation's narrow epistemic merits. In particular, processes can be evaluated along the practical dimension of how they function within our overall cognitive lives. While hitting on a true explanation might be the best way to improve one's narrow epistemic position—i.e., how one thinks about a particular problem or set of propositions—explaining can have more global impact by (for instance) streamlining one's thoughts. Given that actual reasoners only have finite cognitive resources, one way a process could be beneficial to the overall goal of forming useful and true beliefs is if it leads one to the same conclusion with a less demanding process.

Admittedly, having a true explanation can have global effects insofar as it affects our overall store of knowledge and beliefs (see Friedman 1974), but even this is a more limited target of evaluation than how the process of explaining can improve the global process of our thinking and our subsequent behavior. This creates additional logical space for defenses of our inferential practices. As an example of a defense along these lines, consider Kelly's (2007) argument that simplicity can bring one to the truth without "pointing at it" by being the best way to arrive at the truth with the fewest number of "mind changes"—a measure of efficiency. Likewise, explaining might somehow lead one to the truth without drawing one directly to some particular explanation. As should be clear, our aim here is not to offer a particular normative defense of EBI, but rather to suggest how the recognition of EBI changes the philosophical landscape, when it comes to normative justification, relative to IBE.

7 Implications for Education

We have been exploring the idea that EBI is an important means by which our engagement with explanations improves our epistemic standing. If this view is right, what are the implications for education? We would not presume that our preliminary articulation of EBI is positioned to shape pedagogy directly, but we will use this final section of the article to propose some promising directions that we hope might guide future research.

For starters, EBI suggests that when teaching explanations, we should focus not only on their truth, but also on the *processes* by which they are generated and communicated. Indeed, the efficacy of active learning in general, and the power of self-explaining in particular, have been well documented in the educational literature (see Chi and Wylie 2014 for a review of the benefits of "constructive" learning—which includes self-explaining—over most other modes of learning). There is also evidence for the value of training students to engage in high-quality self-explanations. McNamara (2004), for

instance, finds that explicitly teaching and practicing self-explanation techniques can help students with low knowledge about a specific domain gain better comprehension of what a text actually says.

More broadly, Schank (2011) reports tremendous success in teaching students by engaging them in educational programs that stress the processes of thinking over the content of specific domains. Explaining plays a central role in virtually all of the cognitive processes he argues are relevant for academic and real-world success, as his basic model of thinking (e.g., Schank 2011, p. 104) revolves around explaining the failures of our predictions. While results such as those discussed in Sect. 5 do not establish Schank's strong claim that a general skill in explaining is *more important* than domain-specific knowledge, it at least makes plausible the claim that explaining is a crucially important part of learning.

Pedagogically, the upshot of EBI is that we would do well in classroom settings to structure activities so that students are encouraged to engage in high-quality explaining acts. One way to do so would be to provide explicit prompts to explain, as in McNamara (2004). However, there might be more indirect ways to encourage explaining as well. In one study (Chi et al. 2001), novice tutors engaged in explanation as 53 % of their overall "instructional moves," which suggests that having students tutor each other might increase their tendency to engage in explanation. Interestingly, a later study (Chi et al. 2008) also found that merely being a passive recipient of tutors' explanations did *not* improve learning, which again suggests the importance of the explaining *process* beyond merely having the correct explanation. Renkl and Atkinson (2002) explore a variety of ways in which problems and examples can be structured to indirectly encourage self-explanation.

In some ways, our argument for EBI parallels the emphasis on process over product found in some areas of contemporary educational research, such as work discussing final form science mentioned above, as well as work on learning progressions (LPs).[13] Learning progressions are educational paths that get students from one point (a "bottom anchor") to another point (a "top anchor") by fostering an environment that encourages students to progressively revise their conception of the "big ideas" that hold together an entire discipline (e.g., Duncan and Gotwals 2015). As with EBI, learning progressions are important because they emphasize not only which outcome is reached, but also the path that learners take to get there. Moreover, they also emphasize the importance of actively engaging in scientific practices (such as explanation) above and beyond merely acquiring true beliefs.

The similarity between EBI and LP raises the prospect that the two ideas can mutually inform theory development. For example, an obvious question one might have about EBI is how one can tell whether students are making progress, since their reasoning may not result in a correct explanation. The work on LPs suggests several possible approaches. For instance, while one measure of success with learning progressions is their tendency to result in correct explanations (Duncan et al. 2009, p. 667), others involve a student's ability to engage in the general methods of scientific reasoning (ibid.), their "ways of talking" and "strategies for solving problems" (Lehrer and Schauble 2015, p. 433), their "internal coherence" (Hammer and Sikorski 2015), and more generally, their strategies and abilities (Eggert and Bögeholz 2010). These methods of assessment could be applied to EBI as well—we can see if the activity of explaining results in better inferences, for example, regardless of whether those inferences are based on true explanations.

There are several benefits to emphasizing EBI over and above IBE. One is that EBI seems to aid students in the sorts of tasks that go beyond rote memory for the content of explanations (recall, for instance, the measures from Walker et al. 2014)—precisely the

[13] We thank an anonymous reviewer for drawing our attention to this comparison.

kinds of tasks that educators often take to be good indicators of genuine understanding (e.g., Chi et al. 1994). Relatedly, focusing on teaching children to be *good explainers* as opposed to offering *good explanations* arguably teaches a more generalizable skill—one they will be able to apply when asked questions that go beyond studied material (cf. the discussion of "science-as-logic" in Lehrer and Schauble 2006). These speculations are, of course, empirical matters that require further research, but we hope they point to the promise of EBI.

There are also several costs in shifting from IBE to EBI; we believe, however, that these are manageable. One might worry about the lack of clearly defined learning outcomes and measures of progress. However, one possible solution to this problem has already been suggested—the sort of practical inference tasks used with young children could be expanded to older science learners. Another possible solution is also suggested by the LP literature—the goal states (so called "top anchors") used to assess the ultimate efficacy of LPs could be used here. One might also be concerned that the variety of explanatory paths students might take makes designing curricula difficult—again, we can lean on the LP literature on "messy middles" (Gotwals and Songer 2010) to try to draw some practical lessons. As Hammer and Sikorski (2015) argue, learning is an extremely complex phenomenon, and so we should not always expect to be able to model it simply.

The shift from IBE to EBI also represents a shift in how to think about learning. Often—especially outside of educational circles—learning is characterized as the outcome of the right data leading to the right beliefs, with little more. EBI, however, calls this characterization into question. Learning doesn't always require new data (we can sometimes learn, for instance, by deduction or by explaining to ourselves), and we can improve our epistemic standing even when the outcome of engaging in processes like explaining is not the generation of true beliefs. That is, the process of explaining itself can put us in a position to better navigate our world, even when the quality of the explanatory products it ultimately generates leave much to be desired.

Acknowledgments We would like to thank the University of California, Berkeley, the John Templeton Foundation *Varieties of Understanding* project, the McDonnell Scholar Award, and NSF Grant DRL-1056712 (to Tania Lombrozo) for support during the writing of this paper.

References

Achinstein, P. (1983). *The nature of explanation*. New York: Oxford University Press.

Bonawitz, E. B., & Lombrozo, T. (2012). Occam's rattle: Children's use of simplicity and probability to constrain inference. *Developmental Psychology, 48*, 1156.

Bromberger, S. (1966). Why-questions. In R. G. Colodny (Ed.), *Mind and cosmos: Essays in contemporary science and philosophy* (pp. 86–110). Pittsburg: University of Pittsburg Press.

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

Chi, M. T., Roy, M., & Hausmann, R. G. (2008). Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. *Cognitive Science, 32*, 301–341.

Chi, M. T., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science, 25*, 471–533.

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219–243.

Craver, C. F., & Ohio Library and Information Network. (2007). *Explaining the brain*. Oxford: Clarendon Press.

Duncan, R. G., & Gotwals, A. W. (2015). A tale of two progressions: On the benefits of careful comparisons. *Science Education, 99*, 410–416.

Duncan, R. G., Rogat, A. D., & Yarden, A. (2009). A learning progression for deepening students' understandings of modern genetics across the 5th–10th grades. *Journal of Research in Science Teaching, 46*, 655–674.

Duschl, R. A. (1990). *Restructuring science education: The importance of theories and their development*. New York: Teachers College Press.

Edwards, B. J., Williams, J. J., & Lombrozo, T. (2013). Effects of explanation and comparison on category learning. In *Proceedings of the 35th annual conference of the cognitive science society.* Austin, TX: Cognitive Science Society.

Eggert, S., & Bögeholz, S. (2010). Students' use of decision-making strategies with regard to socioscientific issues: An application of the rasch partial credit model. *Science Education, 94*, 230–258.

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

Garfinkel, A. (1980). *Forms of explanation*. New Haven: Yale University Press.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist, 52*(1), 45.

Gopnik, A. (1998). Explanation as orgasm. *Minds and Machines, 8*(1), 101–118.

Gotwals, A. W., & Songer, N. B. (2010). Reasoning up and down a food chain: Using an assessment framework to investigate students' middle knowledge. *Science Education, 94*, 259–281.

Hammer, D., & Sikorski, T. (2015). Implications of complexity for research on learning progressions. *Science Education, 99*, 424–431.

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

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

Hume, D. (2000). *An enquiry concerning human understanding: A critical edition*. Oxford: Oxford University Press.

Jefferys, W. H., & Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist, 80*, 64–72.

Kelly, K. T. (2007). How simplicity helps you find the truth without pointing at it. In V. Harazinov, M. Friend, & N. Goethe (Eds.), *Induction, algorithmic learning theory, and philosophy* (pp. 111–143) Berlin: Springer.

Legare, C., & Lombrozo, T. (2014). The selective benefits of explanation on learning in early childhood. *Journal of Experimental Child Psychology, 126*, 198–212.

Lehrer, R., & Schauble, L. (2006). Scientific thinking and science literacy. In K. A. Renninger, I. E. Sigel, W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology* (Vol. 4, pp. 153–196). Hoboken: Wiley.

Lehrer, R., & Schauble, L. (2015). Learning progressions: The whole world is NOT a stage. *Science Education, 99*, 432–437.

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

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

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

Lombrozo, T. (2012). Explanation and abductive inference. In K. J. Holyoak & R. G. Morrison (Eds.), *Oxford Handbook of Thinking and Reasoning,* (pp. 260–276). Oxford: Oxford University Press.

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes, 38*(1), 1–30.

Papineau, D. (1993). Philosophical naturalism. Oxford: Blackwell.

Pennington, N., & Hastie, R. (1986). Evidence evaluation in complex decision making. *Journal of Personality and Social Psychology, 51*, 242.

Pennington, N., & Hastie, R. (1988). Explanation-based decision making: Effects of memory structure on judgment. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 14*(3), 521.

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.

Persson, J. (2007). IBE and EBI. In J. Persson & P. Ylikoski (Eds.), *Rethinking explanation* (pp. 137–147). Berlin: Springer.

Quine, W. V. O. (1951). *Two dogmas of empiricism. Philosophical Review, 60*(1), 20–43.

Renkl, A., & Atkinson, R. K. (2002). Learning from examples: Fostering self-explanations in computer-based learning environments. *Interactive Learning Environments, 10*(2), 105–119.

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

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

Schank, R. C. (2011). *Teaching minds: How cognitive science can save our schools*. New York: Teachers College Press.

Sobel, D. M., Yoachim, C. M., Gopnik, A., Meltzoff, A. N., & Blumenthal, E. J. (2007). The blicket within: Preschoolers' inferences about insides and causes. *Journal of Cognition and Development, 8*(2), 159–182.

Van Fraassen, B. C. (1980). *The scientific image*. Oxford; New York: Clarendon Press; Oxford University Press.

Walker, C. M., Lombrozo, T., Legare, C. H., & Gopnik, A. (2014). Explaining prompts children to privilege inductively rich properties. *Cognition, 133*, 343–357.

Wilkenfeld, D. A. (2013). Understanding as representation manipulability. *Synthese, 190*(6), 997–1016.

Wilkenfeld, D. A. (2014). Functional explaining: A new approach to the philosophy of explanation. *Synthese, 191*(14), 3367–3391.

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

Williams, J. J., & Lombrozo, T. (2013). Explanation and prior knowledge interact to guide learning. *Cognitive Psychology, 66*(1), 55–84.

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.

Wittgenstein, L., & Anscombe, G. E. M. (2001). *Philosophical investigations: The german text, with a revised English translation [Philosophische Untersuchungen. English and German]* (3rd ed.). Oxford: Blackwell.

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

Woodward, J. (2014). Scientific explanation. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Stanford: Stanford University