Lombrozo, T. (in press). "Learning by Thinking in Science and in Everyday Life." In P. Godfrey-Smith & A. Levy (Ed.), *The Scientific Imagination*. Oxford University Press.

Abstract: This chapter introduces "learning by thinking" (LbT) as a form of learning distinct from familiar forms of learning through observation. When learning by thinking, the learner gains genuinely new insight in the absence of novel observations "outside the head." Scientific thought experiments are canonical examples, but the phenomenon is much more widespread, and includes learning by explaining to oneself, through analogical reasoning, or through mental simulation. The chapter argues that episodes of LbT can be re-expressed as explicit arguments or inferences but are neither psychologically nor epistemically reducible to explicit arguments or inferences, and that this partially explains the novelty of the conclusions reached through LbT. It also introduces a new perspective on the epistemic value of LbT processes as practices with potentially beneficial epistemic consequences, even when the commitments they invoke and the conclusions they immediately deliver are not themselves true.

Keywords: learning, learning by thinking, mental simulation, thought experiments, explanation, self-explanation, psychology, epistemology

Introduction

Two models of learning have dominated both research in human cognition and accounts of scientific progress. The first involves learning from observations, be it everyday experience or the results of systematic research. The second involves learning from testimony, be it the statements of relevant experts or the scientific canon on which new research is based. In both cases, learning is based on new evidence acquired "outside the head."

But some of the time, everyday learning and scientific progress depart from these familiar forms. Consider a child who puzzles through a tricky riddle: What has a single eye but cannot see? When she finally reaches the answer—a needle—she has learned something new. Consider Einstein's well-known thought experiments involving elevators and moving trains, which likewise taught him (and the world) something new. In both cases, the new insight occurred in the absence of novel empirical observations or novel testimony.

I refer to such cases of learning as learning by thinking. Learning by thinking contrasts with the most familiar forms of learning, learning from observation and learning from testimony (itself a special kind of observation). In cases of learning by thinking (LbT), new insight is achieved in the absence of novel observations obtained "outside the head." Such cases naturally raise questions about how such novel insight is possible, in what sense it is really new, and whether we're justified in believing the conclusions delivered by LbT.

In the present chapter, my aim is to review some of what we've learned about LbT from recent research in cognitive development and cognitive psychology and, based on this research, to argue for a new take on the epistemic role of LbT. I'll begin by considering the most widely discussed case of LbT: thought experimentation. The literature on thought experiments within philosophy raises a useful comparison between thought experiments and arguments, which structures the three sections that follow. These sections discuss whether LbT is formally reducible to argumentation (yes), psychologically reducible to argumentation (no), and epistemically reducible to argumentation (no). I ultimately suggest that psychological irreducibility explains the apparent novelty of the conclusions reached through LbT, and I point to a novel take on the epistemic value of LbT processes as practices with potentially beneficial epistemic consequences, even when the commitments they invoke and the conclusions they immediately deliver are not themselves true.

1 Thought Experiments, Arguments, and Three Kinds of Reduction

Thought experiments are canonical examples of learning by thinking. Within philosophy, both scientific and philosophical thought experiments have been the targets of careful analysis, with the challenge being to explain how we seem to learn something new in the absence of novel observations. Articulating this challenge, Kuhn writes: "How, then, relying exclusively upon familiar data, can a thought experiment lead to new knowledge or to a new understanding of nature?" (Kuhn [1964] 1977, p. 241).[1] Working in psychology and education, John Clement asks:

[1] Philosophers vary in how they articulate the puzzle of thought experimentation, some focusing on new knowledge, others on new understanding, etc. For the most part, the discussion has not focused on new learning, which is the main focus in the discussion that follows. I am grateful to Mike Stuart for bringing this point to my attention.

"How can findings that carry conviction result from a new experiment conducted entirely within the head?" (Clement 2009, p. 687).

One approach is to reduce thought experiments to more familiar forms of learning. For instance, John Norton argues that thought experiments are truly arguments, perhaps disguised in picturesque, narrative form (Norton 1996). On this view, thought experiments generate something "new" in the sense that they derive a novel conclusion from known premises by applying deductive or inductive rules of inference.[2] A nice feature of this view is that the conclusions delivered through thought experimentation can potentially be justified—this will transpire just in case (and to the extent that) the corresponding argument justifies its conclusion.

Thought experiments might also share properties with learning through observation. For example, Mach suggested that thought experiments reflect "instinctive knowledge" gathered through experience (Mach 1897, 1905)—a body of implicit (but potentially justified) beliefs that can be accessed through thought experimentation, yielding novel insights directly or as premises in further arguments. Within psychology, Clement suggests that mental simulations can "draw out implicit knowledge" contained in mental schemata that "the subject has not attended to and/or not described linguistically before" (2009, 694). There are a variety of alternative proposals, including some with more rationalist (e.g., Brown 1991) and evolutionary commitments (e.g., Shepard 2008).

For present purposes, the comparison between thought experiments and arguments is useful in framing a set of related questions about whether—and in what sense—learning by thinking is reducible to argumentation. In a paper on thought experiments in science, for example, Tamar Gendler asks whether "any conclusion reached by a good thought experiment will also be demonstrable by a non-thought-experiment argument" (1998, 399), and she goes on to differentiate three readings of "demonstrable" that correspond to three questions about reducibility. Specifically, she asks whether thought experiments can be reconstructed as arguments from the perspective of a mature science, whether they nonetheless have heuristic value, and whether they are epistemically equivalent within a developing science. Her answers are, respectively, (trivially) yes, (trivially) yes, and (controversially) no.
In the present chapter, I take up a related set of questions.

First, are thought experiments formally reducible to arguments?[3] That is, is there some argument, with appropriate premises, rules of inference, and conclusions, that delivers the conclusions of the thought experiment? Second, are thought experiments psychologically reducible to arguments, in the sense that any conclusions reached through thought experimentation by a given person could, under the same circumstances, have also been reached through an explicit argument? And finally, are thought experiments epistemically reducible to arguments, in the sense that the conclusions of thought experiments derive their epistemic force entirely and exclusively from the force of the corresponding argument? My answers (yes, no, and no) will mirror Gendler's, but my analysis

[2] There's still something interesting to be said about the sense in which deduction (or induction) generates something "new." For relevant discussion, see Powers 1978.

[3] It is worth clarifying that the notion of "argument" used throughout the chapter is deliberately broad. I include as arguments any inferences that can be represented in terms of premises, conclusions, and rules of inference, even if the premises or rules of inference are not ones we would typically offer in a verbal argument. For example, an application of Bayes's rule could feature in an argument. This notion is therefore broader than that employed, for instance, in the argumentative theory of reasoning (Mercier and Sperber 2011). See also Stuart 2016.

differs from hers in two important ways: in diagnosing what it is that makes the conclusion of a thought experiment "new," and in differentiating the epistemic roles of thought experiments and arguments.

These questions, while formulated here in terms of thought experimentation, arise for any LbT process. In the three sections that follow, I consider each question in turn, relying most heavily on research that involves learning by explaining to oneself.

2 The Case for Formal Reduction

Within psychology, there has been little research on thought experimentation as such (for exceptions, see Clement 2009). However, psychologists have studied mental simulations (Hegarty 2004), which are very much like thought experiments, as well as other processes that involve LbT, such as explaining to oneself (Fonseca and Chi 2011; Lombrozo 2012, 2016) and engaging in analogical reasoning (Gentner and Smith 2012). Based on this work, my colleagues and I have argued that LbT processes effectively recruit constraints on reasoning that deliver conclusions that might not otherwise be reached (Lombrozo 2012, 2016), where these constraints play a role akin to premises or rules of inference within an argument.

To better appreciate the basis for these ideas, consider a typical experiment involving learning by explaining. In such experiments, participants are presented with a task, such as learning to categorize novel objects or learning what activates a machine. Half the participants are prompted to explain to themselves at key points in the experiment. For example, they might be asked to explain why a particular object belongs to a particular category, or to explain why a particular object activated a machine. Importantly, participants never receive feedback on the content or quality of their explanations. In the control condition, participants are instead asked to engage in a task that's comparably demanding, such as thinking aloud, describing category members, or reporting whether or not a given object activated the machine. Participants are then probed to assess whether those who explained differ from those in the control condition in terms of the inferences they draw or the information they recall. If the former group outperforms the latter, this constitutes evidence for LbT, as participants were all presented with the same evidence and the same probes; the differences can be attributed to the kind of thinking in which they engaged.[4]

Using experiments that follow this basic form, we have found that relative to participants in control conditions, both children and adults who are prompted to explain are more likely to discover and to generalize patterns that support broad and simple explanations (Kon & Lombrozo in press, Walker, Bonawitz, and Lombrozo 2017, Walker and Lombrozo 2017, Walker et al. 2014, 2016, 2017; Williams and Lombrozo 2010, 2013; Williams et al. 2013) and to privilege causal information over superficial perceptual properties (Legare and Lombrozo 2014; Walker et al. 2014). For example, in one study, participants studied eight novel robots, four of which

[4] One might worry that requesting an explanation is itself a kind of evidence. For example, it might bring with it the implication that there is something that can easily be explained, such that the experimenter expects the participant to have discovered it. Various experiments have aimed to equate such pragmatic inferences across conditions (Williams et al. 2013) or to determine whether participants do draw such inferences (Williams and Lombrozo 2010). The research to date suggests that effects of explanation cannot be attributed to these factors.

belonged to one category and the remaining four to another (Williams and Lombrozo 2010). The examples were constructed such that the two groups of robots could be differentiated by a salient but imperfect rule: three of the four robots in one category had round bodies, and three of the four robots in the other category had square bodies. The two groups could also be classified perfectly by discovering and using a more subtle rule: all of the robots in one group had feet that were flat on the bottom, and the remaining four had feet that were pointy on the bottom (despite otherwise variable foot shapes). Participants who were asked to explain why each robot might belong to its respective category were significantly more likely to discover this more subtle basis for classification, and to use it subsequently in classifying novel robots.

Why might explaining have these effects? Williams and Lombrozo (2010, 2013) propose what they call the "subsumptive constraints" account, according to which the process of explaining recruits an important explanatory constraint: to identify an explanans that explicitly or implicitly invokes an explanatory generalization that subsumes the explanandum. In so doing, participants will be driven to seek and favor broad patterns—that is, those that they believe apply to many cases—over idiosyncratic ones. This makes the potentially counterintuitive prediction that prompts to explain might actually impair learning when there is no broad pattern to be found, and indeed, this is what we've found (Williams et al. 2013).

The subsumptive constraints account is one piece of a larger story about explanation (Lombrozo 2011, 2012, 2016) that has affinities to "inference to the best explanation" (IBE) in philosophy (Harman 1965; Lipton 2003). In the case of IBE, the core idea is that explanatory virtues—such as scope and simplicity—can inform an inference to which explanation is true. When it comes to our account of learning by explaining, the core idea is that the process of engaging in explanation recruits explanatory virtues as evaluative criteria, and these in turn act as constraints on learning and inference by leading learners to seek and privilege hypotheses that support those virtues. In the language of argument structure, the explanatory virtues are like premises or rules of inference (inductive constraints) that favor some conclusions over others.[5]

To make these claims more concrete, it helps to consider another example, this time drawn from work with five-year-old children (Walker, Bonawitz, and Lombrozo, 2017). We know from prior work that adults favor explanations for two effects that are "simple" in the sense that they appeal to a common cause over those that appeal to two independent causes (Lombrozo 2007), and that this is driven by a preference for explanations that invoke the fewest unexplained causes, not the fewest causes per se (Pacer and Lombrozo 2017). The preference for common-cause over independent-cause explanations has also been found for preschool-aged children (Bonawitz and Lombrozo 2012). If the process of engaging in explanation recruits explanatory virtues such as simplicity, then we should expect to see a greater role for simplicity as a constraint on inference when children engage in explanation than when they do not.

[5] One can potentially align explanatory virtues (such as a preference for broad scope or greater simplicity) with either premises or rules of inference, and either approach is consistent with the data reported here. Determining which kind of process or representation in fact governs human behavior suffers from especially acute problems of underdetermination (Anderson 1978). In part for this reason, I often refer to explanatory virtues as "constraints" on learning and inference, as this locution is neutral with respect to the underlying representation or process.

Walker and colleagues tested this prediction by presenting five-year-old children with an illustrated garden from which carrots could be sampled, revealing which were healthy and which were "sick." Children initially saw two sick carrots, one sampled after the other, and were asked either to explain why the plants were sick (i.e., "Why do you think these plants are sick?") or, in a control condition, to report what they observed (i.e., "Were these plants healthy or sick?"). Crucially, these observations were consistent with two explanations: one appealing to a common cause (both were sick because they were in the area with red soil) and the other to two plausible but independent causes (one was sick because it was in the shade of a tree, another was sick because it was near a broken sprinkler). The five-year-olds who were prompted to explain were significantly more likely than those in the control condition to make subsequent inferences in line with the simple explanation.[6] It appears that engaging in explanation increased the extent to which they recruited simplicity as an inductive constraint, and that this accounts for the effects of "mere thinking" on learning.

In sum, research with both children and adults has documented systematic effects of engaging in explanation on learning and inference—even in the absence of feedback on the accuracy or quality of explanations. This form of self-explaining is an instance of LbT that, like thought experimentation, occurs in the absence of evidence obtained outside the head. While effects of explanation on learning are almost certainly driven by multiple mechanisms, the research highlighted here points to one particular facet of learning by explaining with close parallels to IBE: the idea that engaging in explanation recruits inferential constraints (namely, scope, simplicity, and other explanatory virtues) that affect subsequent learning and reasoning. If this account is right, learning by explaining is formally reducible to a kind of argumentation, with explanatory constraints featuring as premises or implicitly in inferential rules.

3 The Case Against Psychological Reduction

The research reviewed in the previous section suggests that the consequences of learning by explaining can be modeled as an inferential process that weights explanatory considerations—such as scope and simplicity—more heavily than they're weighted when engaged in other processes, such as passively observing or thinking aloud. This naturally raises the question of why explaining is necessary to reach particular conclusions. That is, are LbT processes "psychologically dispensable" in the sense that they can readily be replaced by alternative forms of reasoning, such as explicit argumentation?

The answer seems to be no. Most generally, LbT processes are uniquely powerful precisely because they deliver conclusions that appeal to premises or inferential rules that are not otherwise available. In her discussion of Galileo's famous thought experiment involving falling bodies, Gendler (1998) suggests that engaging in a mental simulation brings in implicit commitments concerning which properties are physically determined. Endorsing aspects of Mach's view, she writes:

[6] The study tested four-year-olds and six-year-olds as well. However, the four-year-olds responded at chance, while the six-year-olds tended to draw inferences in line with the simpler explanation regardless of whether they were prompted to explain. While these developmental changes are interesting in their own right, and discussed in Walker, Bonawtitz, and Lombrozo 2017, they are not relevant to the point made here.

> We have stores of unarticulated knowledge of the world which is not organized under any theoretical framework. Argument will not give us access to that knowledge, because the knowledge is not propositionally available. Framed properly, however, a thought experiment can tap into it, and—much like an ordinary experiment—allow us to make use of information about the world which was, in some sense, there all along, if only we had known how to systematize it into patterns of which we were able to make sense. (Gendler 1998, 415)

Based on analyses of scientifically trained experts reasoning aloud through novel problems, Clement relatedly suggests that mental simulations begin from "implicit physical intuitions apprehended via imagistic simulations, rather than explicit linguistic propositions or axioms" (2009, 704). Because the bases for thought experimentation need not be represented linguistically, they may not be accessible via other forms of reasoning, such as explicit argumentation (see also Miscevic 1992, Nersessian 2007).

These views rest on substantive—but plausible—commitments concerning human cognitive architecture. In particular, they rest on the idea that different mental representations are available to different mental processes. Linguistic representations may be available to the processes involved in explicit argumentation, while other forms of mental content—such as perceptual and motor memories, or explanatory virtues—may only emerge as constraints on reasoning when a thinker is engaged, respectively, in mental simulation or explanation.[7] In support of these claims, consider two examples: one involving motor and perceptual simulation, the other explanatory virtues.

In a 1999 paper, Schwartz and Black report an experiment in which participants were invited to imagine a narrow cylindrical cup and a wide cylindrical cup of equal heights, each filled with water to the same height. Participants were asked what would happen as the two cups are tilted: would they begin to pour water when tilted to the same angle, or at different angles? And if they would pour at different angles, which would require a greater tilt? When asked explicitly, a minority of participants (18.8%) gave the correct answer: that the narrow cup would need to be tilted farther. When asked to actually tilt the cups, with eyes closed, to the point at which imaginary water would begin to pour, 100% of participants correctly tilted the narrower glass to a greater degree.[8] In a subsequent experiment that involved visualizing this motion—without actually holding a glass or moving one's hands—participants were again more accurate than their explicit judgments. This study

[7] Note that this is a more radical form of pluralism than that endorsed by popular "two-systems" approaches within psychology (e.g., Evans and Stanovich 2013; Kahneman and Frederick 2002; Sloman 1996), in that many representational formats and processes must be differentiated. However, this form of pluralism need not take on the additional commitments associated with dual systems approaches, e.g., that systems are either automatic or controlled. Moreover, the effects of engaging in LbT processes, such as explanation, should not be equated with a shift from System 1 to System 2. In most of the experiments on explanation reported here, explanation is contrasted with a control task that is similarly deliberative and that requires the use of language. On most taxonomies, both the explanation condition and the control condition fall on the more controlled and deliberative side of the dichotomy.

[8] Schwartz and Black (1999) conducted three versions of this task using differently shaped cups: rectangular, cylindrical, and cone-shaped. The numbers reported here correspond to performance with the cylindrical cup. In all three cases, participants' explicit judgments were considerably less accurate than their tilting behavior.

provides evidence that motor and perceptual simulations can offer information that isn't otherwise available to inform judgments.

As a second example, consider a finding from Pacer and Lombrozo (2017). In a series of experiments, participants were asked to provide the most satisfying explanation for an alien's two symptoms, where the viable options contrasted two plausible metrics for simplicity in causal explanations: "node" simplicity, according to which the simpler explanation is the one that invokes fewer causes, and "root" simplicity, according to which the simpler explanation is the one that invokes fewer unexplained causes. Participants reliably chose explanations that were lower in root simplicity but not node simplicity, and treated root simplicity as a virtue commensurate with probabilistic information. Yet when asked to justify their explanation choices, participants almost never appealed to a notion like simplicity or parsimony, and never identified the virtue that seemed to actually guide judgments: reducing the number of unexplained causes. This suggests that the explanatory constraint invoked through explanation—in this case a preference for low root simplicity—was not available to explicit reason in a way that would likely inform explicit argumentation or other explicit forms of inference.

These examples support the psychological commitments implicit in proposals by Mach, Gendler, and others. They also suggest a modest sense in which LbT processes, such as mental simulation and self-explanation, can offer something new: they create a representation with novel affordances, one that's newly available to processes of explicit reasoning and argumentation, no matter that in some sense the relevant knowledge was there all along.[9]

4 The Case Against Epistemic Reduction

So far I've argued for a sense in which learning by thinking is formally reducible to argumentation, but that psychological reality is such that LbT processes can sometimes deliver conclusions that could not have been reached through explicit argumentation. In brief, LbT processes provide access to constraints on learning and inference—what can be thought of as premises or inference rules—that are not available through explicit argumentation. These aspects of the paper correspond, respectively, to the case for formal reduction and against psychological reduction. We can now turn to epistemic reduction. That is, do the conclusions delivered through LbT processes have the same epistemic status as those of the corresponding formal arguments?

Philosophers have debated this question for the case of thought experiments. Advocates of the "argument view," such as Norton (1996), naturally endorse epistemic reduction. Good thought experiments correspond to good arguments, bad thought experiments to bad arguments. A thought experiment is precisely as epistemically powerful as its corresponding argument. Others, such as Gendler (1998), argue that some epistemic force is lost in translation. For Gendler, this is in

[9] Gendler argues for a stronger sense in which scientific thought experiments can yield something new. Regarding Galileo's thought experiment involving falling bodies, she writes: "The thought experiment that Galileo presents leads the Aristotelian to a reconfiguration of his conceptual commitments of a kind that lets him see familiar phenomena in a new way. What the Galilean does is provide the Aristotelian with conceptual space for a new notion of the kind of thing natural speed might be: an independently ascertainable constant rather than a function of something more primitive (that is, rather than a function of weight). It is in this way, by allowing the Aristotelian to make sense of a previously incomprehensible concept, that the thought experiment has led him to a belief that is properly taken as new" (1998, 412).

part because psychological reduction fails. She writes: “Even if it could be replaced with an equally effective argument, the justificatory force of a thought experiment might still be based on its capacity to make available in a theoretical way those tacit practical commitments which enable us to negotiate the physical world” (1998, 415). To the extent those tacit commitments are themselves justified, they offer some justificatory force we can’t otherwise achieve, as it’s the very process of thought experimentation that makes those tacit commitments available.

Although the two perspectives just described differ considerably, they share a basic assumption. On both views, thought experiments are justified to the extent the commitments they invoke (however implicitly) are justified—and, presumably, to the extent the inference rules they use are truth-preserving. This seems intuitive enough, but it isn’t the only way to approach the epistemic value of LbT processes. In particular, LbT processes could potentially yield justified conclusions even when the premises they invoke are false. More radical still, engaging in certain forms of LbT could be epistemically beneficial (in the sense that they foster justified beliefs as a downstream consequences) even when the immediate conclusions they deliver are false.

As a candidate instance of this first possibility, consider an account of thought experiments offered by Hayley Clatterbuck (2013). Clatterbuck’s account rests on a type of inductive inference that she calls “Dewey induction,” following a distinction articulated by Peter Godfrey-Smith (2011). In Dewey inductions, generalizations from known to unknown cases derive their force not from the statistical properties of a sample (for instance, that it is large and that sampling was random) but from the characteristics of the known case: it must be representative of its kind. Clatterbuck writes that some thought experiments are “instances par exemplar of Dewey inductions,” where their force derives from their “ability to generate an inductive argument that does not depend on enumerative induction” (2013, 320). To generate thought experiments that support Dewey inductions, the reasoner first simulates a phenomenon known from experience, and then idealizes the case to remove contingent details, thereby (in a successful thought experiment) rendering it representative of its kind, and a good basis for generalizing to novel instances. The idealization step is central to Clatterbuck’s argument, and it also provides the crucial link to the point I aim to make here, as idealizations, in an important sense, are fictions.[10] If Clatterbuck’s account is right, then thought experiments can sometimes yield justified conclusions, even though some of the commitments they import depend on a process of idealization that deliberately distorts what we’ve actually observed from direct experience. Their epistemic value might not derive—at least not directly—from the truth of implicit commitments they invoke.

Consider now the more radical possibility alluded to before: that in some cases LbT could be epistemically beneficial not only when the commitments invoked are false but also when the immediate conclusion supplied is false. To do so, let’s return to the case of simplicity in explanation. One justification for favoring simpler explanations comes from Newton, who writes, in the Principia Mathematica, that “we are to admit no more causes of natural things than such as are both true and

[10] It’s not clear whether Clatterbuck herself takes idealizations to be fictions. Her paper assumes that some idealizations can be “better” than others, but this kind of evaluation is consistent with the view that idealizations deliberately mis-describe the world. She certainly does suggest that idealization involves removing information from experientially familiar cases. Others who take idealizations or scientific models to incorporate fictional elements include Frigg 2010, Godfrey-Smith 2009, Levy 2015, and Toon 2010.

sufficient to explain their appearances . . . for Nature is pleased with simplicity, and affects not the pomp of superfluous causes" (Newton [1687] 1964, p. 398). In other words, we should favor explanations that involve fewer causes, and this is justified because the world is itself simple. The constraint invoked through explanation—effectively, "simpler is more likely"—is epistemically warranted (so Newton seems to imply) because it is true. This defense of simplicity is consistent with epistemic reduction: the justification for an LbT conclusion derives from the justification for the premises (implicitly) invoked, as in an argument.

Contrast this approach to simplicity with that developed by Kevin Kelly (2007). Kelly formalizes a different metric for simplicity, and he demonstrates that under appropriate assumptions, favoring simpler hypotheses will lead to the right answer with a smaller number of mind-changes. On this view, there's epistemic value to favoring simplicity: it gets us to true beliefs more efficiently. But insofar as there's an epistemic justification for favoring simplicity, it doesn't require an assumption that simpler hypotheses are more likely. Instead, the benefits are further downstream: favoring simplicity helps us get to true beliefs . . . eventually. A psychological mechanism that implements this process could therefore guide us to true beliefs, even though the commitments embedded in the inferential process that generates those beliefs—"simpler is better"—need not be themselves "true" in the sense that they directly describe or resemble the world, and even though the outcome of favoring simpler explanations will often be a false (but temporary) belief.

Clatterbuck's and Kelly's positions help sketch out the possibility that LbT processes could have positive epistemic consequences even when the premises they invoke are false, and even when the conclusions they deliver (at least in the short term) are false. The proposal has some empirical support as well. Here, again, the most compelling evidence comes from the case of learning by explaining. In many cases, learning by explaining has beneficial effects because the constraints invoked through explanation accurately mirror the structure of what's being learned (e.g., Williams et al. 2012). Explaining thus helps a learner arrive at the correct explanation, and having the correct explanation accounts for many of the beneficial consequences of having engaged in explanation. But in some cases there are benefits to engaging in explanation even when the explainer fails to generate an explanation, or generates an explanation that is false. How could this be?

One example of this phenomenon comes from research reported by Chi et al. (1994). In their experiment, eighth-grade students studied a text about the human circulatory system, with some students prompted to explain to themselves (without feedback) after each line of the text and others prompted to read through the materials twice. The researchers documented learning benefits for those prompted to explain, even though the explanations were often incorrect. They suggest that generating an explanation "objectifies" the incorrect commitments it embodies in a way that allows learners to recognize a conflict between those commitments and the accurate text they're simultaneously reading. Recognizing the conflict can, in turn, initiate a process of belief revision.

Interestingly, this proposal seems to presuppose a kind of psychological irreducibility, as the commitment that conflicts with the text becomes available for scrutiny (and rejection) when a learner engages in explanation, but not when a learner engages in a control task. It also shares characteristics with accounts of "destructive" thought experiments, which help render inconsistencies apparent (e.g., Brown 1991. The critical point for our discussion of epistemic irreducibility is this: the benefits of engaging in an LbT process need not derive from the immediate

conclusion that the LbT process renders available (the correct or incorrect explanation). Epistemic benefits can also occur as downstream consequences, in this case a metacognitive awareness of inconsistency that triggers belief revision, eventually leading to more accurate beliefs.

As a second example, consider findings from Walker et al. (2014). In their first study, three- to five-year-old children were presented with sets of three blocks, where a target block in each set had a causal property (it made a toy play music when placed on top of it) and a perceptual property (e.g., a yellow exterior). The remaining two blocks each shared one property with the target: the "causal match" made the toy play music but was a different color; the "perceptual match" was the same color but did not make the toy play music. Children saw each block go on the toy, with half the children prompted to explain why the block did (or did not) make the toy play music, and the other half, in a control condition, prompted to report whether the block did (or did not) make the toy play music.

After all three blocks had been placed on the toy, one after another, the experimenter revealed that the target block had a hidden internal part (a red pin). Children were asked to indicate which of the other blocks—the causal match or the perceptual match—was more likely to share the internal part. Replicating prior work (Sobel et al. 2007), the study showed that the older children were more likely than younger children to generalize the internal part to the causal match over the perceptual match. In addition, however, those children who had been prompted to explain were significantly more likely than those in the control condition to generalize to the causal match over the perceptual match.

Here's one account of these results. When asked to explain why blocks did or did not make the toy play music, children were more likely to posit unobserved causal mechanisms, and therefore to expect similarities in internal structure that tracked causal affordances. In fact, many children did generate explanations that appealed to internal parts or mechanisms (e.g., "because it has something inside of it"; "because it has batteries"), and children who generated such explanations were more likely than those in the control condition to generalize the internal part on the basis of causal rather than perceptual similarity. But even children who produced other kinds of explanation—such as those appealing to appearance ("because it's purple") or kind membership ("because it's a music-maker")—were more likely than children in the control condition to generalize on the basis of causal over perceptual similarity.

What was explanation doing in such cases? It seemed to generate a more "adult-like" pattern of generalization, no matter that the explanations themselves didn't point to internal parts. Wilkenfeld and Lombrozo (2015) identify a variety of mechanisms that could be operating in such cases. Beyond the broadly metacognitive benefits suggested by Chi and colleagues, explaining could engage other processes that have positive downstream consequences, such as comparison (Edwards et al. 2019) and abstraction (Walker and Lombrozo 2017, Walker et al. 2014; Williams and Lombrozo 2010), both of which facilitate the extraction and application of rules and general schemata (Gentner and Medina 1998). These processes could in turn affect reasoning, even if the immediate output of the LbT process—the explanation—is not itself veridical or the basis for an appropriate inference.

Wilkenfeld and Lombrozo coin the term "explaining for the best inference" (EBI) in characterizing a practice that encompasses such cases. Unlike inference to the best explanation (IBE), EBI focuses on the downstream consequences of

engaging in explanation, not the immediate inferential consequences of privileging particular explanations. EBI therefore suggests a kind of epistemic question different from that traditionally posed in the case of thought experimentation. Rather than focusing on whether the conclusions delivered by LbT processes are justified, where their justification derives from the epistemic status of the premises and inference rules involved in their generation, we can instead ask whether the practice of engaging in LbT processes is, on the whole, epistemically valuable in the sense that, downstream, it leads us to a better suite of beliefs.

The shift from thinking about the epistemic status of LbT commitments to LbT practices has parallels in the literature on modeling in science. Specifically, Levy (2012) introduces a useful distinction between two approaches to scientific models. His aim is to explain how models can be fictions while operating with realist commitments. Toward this end, he introduces "indirect realism" and "modeling as metaphor."

Levy's first option—indirect realism—holds that scientific models should be understood as wholly fictional: the entities and relations they posit are imaginary, not real. The model is thus an object of scientific study in its own right, but comparing the model to the system it targets offers "a way of converting knowledge about the model to knowledge about the world" (2012, 742). For instance, one might regard models as sharing a similarity relation to their targets (e.g., Weisberg 2012, such that we can generalize features of the model to the world when the appropriate similarity relations obtain.

On Levy's second option, "modeling as metaphor," models aren't wholly fictional: they are about real entities and relations. However, we know that models often simplify and idealize target systems: they deliberately "mis-describe," and are in this weaker sense fictional. The interesting move comes in reconciling this approach to modeling with a form of realism. Levy suggests that rather than regarding the aim of a realist picture of science to be the production of "true" theories and models, we can shift to a picture in which the aim is the production of true beliefs. He writes:

In most formulations of realism the locus of the doctrine is seen as the content of the theory or model. The view is that scientists aim to attain true models. But we might also view realism as a doctrine concerning true beliefs. The idea would be, roughly, that realism is the doctrine that science aims to allow us to acquire knowledge about the world. . . . [I]f realism is a doctrine about knowledge, then theoretical science can be successful, from the realist's point of view, even if its immediate products, e.g. models, are false. Deliberate distortions of the truth are fine, so long as models allow us to form (and justify) correct beliefs about the world. (2012, 743)

In other words, we can shift from thinking about models as epistemically valuable to the extent they accurately describe or approximately resemble the world to instead considering their epistemic value in terms of their role in supporting the acquisition of true beliefs. A model can be false, but a downstream consequence of engaging in the process of modeling can be the production of true beliefs.

Not all instances of scientific modeling involve LbT: models are often updated in light of observations "outside the head," and they're often employed in simulations implemented on computers, not human minds. Nonetheless, learning from models and learning by thinking share obvious parallels, and focusing on cases where these practices are beneficial—despite fiction, idealization, or inaccuracy—makes Levy's suggested account of realism attractive for the account of LbT sketched here. Just

as “metaphorical” models can play a role in scientific progress, LbT processes might improve our epistemic potential, even when the commitments they invoke and the conclusions they deliver aren’t strictly true. Instead, the practice of engaging in LbT might “allow us to form (and justify) correct beliefs about the world.”

In this section, I’ve sketched a view according to which LbT processes are not epistemically reducible to arguments. Specifically, engaging in LbT can have positive downstream epistemic consequences, but because LbT processes are not psychologically reducible to argumentation, these consequences will not, as a rule, be achieved by substituting LbT for explicit argumentation. For example, engaging in explanation seems to promote comparison and abstraction, and benefits learners even when they fail to arrive at an accurate explanation (Wilkenfeld and Lombrozo 2015); it’s doubtful that the corresponding arguments would generate the same effects. What I haven’t done is show that LbT processes are guaranteed or even likely to have positive effects. In part this is because each LbT process recruits a unique set of constraints, and each will correspondingly require a custom argument for why those constraints will tend to yield particular epistemic consequences in particular contexts. Developing and testing such accounts is beyond the scope of this chapter, but are important directions for future work.

5 Conclusions

Learning by thinking is pervasive in science and in everyday life. While the most celebrated examples—such as Galileo’s and Einstein’s thought experiments—are rare indeed, their more mundane counterparts, including mental modeling and simulation, explaining to oneself, and engaging in analogical reasoning, occur on a regular basis. Drawing on philosophical work on thought experimentation and empirical work on learning by explaining, I’ve suggested answers to three questions about the reducibility of LbT to argumentation. Specifically, I’ve argued that LbT processes are formally reducible to their corresponding arguments, but that they are neither psychologically nor epistemically reducible to their corresponding arguments.

The case against psychological reduction offers a modest sense in which LbT processes offer something “new”: they make commitments available to new cognitive processes, such as explicit verbal reasoning. The case against epistemic reduction, while more tentative, offers a new way of approaching the epistemic value of LbT practices. Rather than focusing on whether particular commitments or conclusions are warranted, we can consider whether particular practices are warranted by virtue of their downstream consequences. Further developing and testing this proposal will surely require more thinking and more argumentation—with observations generated both inside and outside the head.

Acknowledgments
This work was supported by a McDonnell Scholar Award in Understanding Human Cognition, as well as NSF grant DRL-1056712. I am also grateful to Peter Godfrey-Smith and Arnon Levy for helpful comments on a draft from September 2015, and to Mike Stuart for helpful conversation and comments in 2016.

References

Anderson, J. R. (1978). “Arguments Concerning Representations for Mental Imagery.” Psychological Review 85: 249–277.

Bonawitz, E. B., and Lombrozo, T. (2012). “Occam’s Rattle: Children’s Use of Simplicity and Probability to Constrain Inference.” Developmental Psychology 48: 1156–1164.
Brown, J. R. (1991). Laboratory of the Mind: Thought Experiments in the Natural Sciences. 2nd ed. London: Routledge.
Brown, J. R., and Fehige, Y. (2014). “Thought Experiments.” The Stanford Encyclopedia of Philosophy (Fall 2014 ed.), edited by E N. Zalta. http://plato.stanford.edu/archives/fall2014/entries/thought-experiment.
Chi, M.T.H., De Leeuw, N. Chiu, M. and LaVancher, C. (1994). "Eliciting self-explanations improves understanding." Cognitive science 18, no. 3: 439-477.
Clatterbuck, H. (2013). “The Epistemology of Thought Experiments: A Non-eliminativist, Non-Platonic Account.” European Journal for Philosophy of Science 3, ino. 3: 309–329.
Clement, J. J. (2009). “The Role of Imagistic Simulation in Scientific Thought Experiments.” Topics in Cognitive Science 1, no. 4: 686–710.
Edwards, B. J., Williams, J. J., Gentner, D., & Lombrozo, T. (2019). Explanation Recruits Comparison in a Category-Learning Task. Cognition , 185: 21-38.
Evans, J. St. B. T., and Stanovich, K. E. (2013). “Dual-Process Theories of Higher Cognition: Advancing the Debate.” Perspectives on Psychological Science 8, no. 3: 223–241.
Fonseca, B. A., and Chi, M. T. (2011). “Instruction Based on Self-Explanation.” Handbook of Research on Learning and Instruction, edited by R. E. Mayer and P. A. Alexander, 296–321. New York: Routledge.
Frigg, R. (2010). “Models and Fiction.” Synthese 172, no. 2: 251–268.
Gendler, T. S. (1998). “Galileo and the Indispensability of Scientific Thought Experiment.” British Journal for the Philosophy of Science 1998: 397–424.
Gentner, D., and Medina, J. (1998). “Similarity and the Development of Rules.” Cognition 65, no. 2: 263–297.
Gentner, D., and Smith, L. (2012). “Analogical Reasoning.” Encyclopedia of Human Behavior, edited by V. S. Ramachandran, 1: 130–136. Oxford: Elsevier.
Godfrey-Smith, P. (2009). “Models and Fictions in Science.” Philosophical Studies 143, no. 1: 101–116.
Godfrey-Smith, P. (2011). “Induction, Samples, and Kinds.” In Carving Nature at Its Joints: Natural Kinds in Metaphysics and Science, edited by M. H. Slater, M. O’Rourke, and J. K. Campbell, 33–52. Cambridge, MA: MIT Press.
Harman, G. H. (1965). “The Inference to the Best Explanation.” Philosophical Review 74, no. 1: 88–95.
Hegarty, M. (2004). “Mechanical Reasoning by Mental Simulation.” Trends in Cognitive Sciences 8, no. 6: 280–285.
Kahneman, D., and Frederick, S. (2002). “Representativeness Revisited: Attribute Substitution in Intuitive Judgment.” In Heuristics and Biases: The Psychology of Intuitive Judgment, edited by T. Gilovich, D. W. Griffin, and D. Kahneman, 49–81. Cambridge: Cambridge University Press.
Kelly, K. T. (2007). “A New Solution to the Puzzle of Simplicity.” Philosophy of Science 74, no. 5: 561–573.
Kon., E. & Lombrozo, T. (in press). Scientific discovery and the human drive to explain. In Richard Samuels & Daniel Wilkenfeld (Eds.), Advances in Experimental Philosophy of Science. New York, NY: Bloomsbury Press.

Kuhn, Thomas. ([1964] 1977). “A Function for Thought Experiments.” In The Essential Tension: Selected Studies in Scientific Tradition and Change. Chicago: University of Chicago Press.
Legare, C. H., and Lombrozo, T. (2014). “Selective Effects of Explanation on Learning During Early Childhood.” Journal of Experimental Child Psychology 126: 198–212.
Levy, A. (2012). “Models, Fictions, and Realism: Two Packages.” Philosophy of Science 79, no. 5: 738–748.
Levy, A. (2015). “Modeling Without Models.” Philosophical Studies 172, no. 3: 781–798.
Lipton, P. (2003). Inference to the Best Explanation. London: Routledge.
Lombrozo, T. (2007). “Simplicity and Probability in Causal Explanation.” Cognitive Psychology 55: 232–257.
Lombrozo, T. (2011). “The Instrumental Value of Explanations.” Philosophy Compass 6, no. 8: 539–551.
Lombrozo, T. (2012). “Explanation and Abductive Inference.” In Oxford Handbook of Thinking and Reasoning, edited by K. J. Holyoak and R. G. Morrison, 260–276. Oxford: Oxford University Press.
Lombrozo, T. (2016). “Explanatory Preferences Shape Learning and Inference.” Trends in Cognitive Sciences, 20, no. 10: 748-759..
Lombrozo, T., and Walker, C. M. (n.d.). “Learning by Thinking.” Manuscript.
Mach, E. (1897). “Über Gedankenexperimente.” Zeitschrift für den physikalischen und chemischen Unterricht 10: 1–5.
Mach, E. (1905). “Über Gedankenexperimente.” In Erkenntnis und Irrtum, 181–197. Leipzig: Johann Ambrosius Barth, 181–197. Translated by J. McCormack, in Knowledge and Error, 134–147. Dordrecht: Reidel.
Mercier, H., and Sperber, D. (2011). “Why Do Humans Reason? Arguments for an Argumentative Theory.” Behavioral and Brain Sciences 34, no. 2: 57–74.
Miščević, N. (1992). Mental models and thought experiments. International Studies in the Philosophy of Science, 6, no. 3: 215-226.
Nersessian, N. J. (2007). Thought experimenting as mental modeling: Empiricism without logic. Croatian Journal of Philosophy, 7, no. 20: 125-161.
Newton, Isaac. ([1687] 1964). The Mathematical Principles of Natural Philosophy (Principia Mathematica), New York: Citadel Press.
Norton, J. D. (1996). “Are Thought Experiments Just What You Thought?” Canadian Journal of Philosophy 26, no. 3: 333–366.
Pacer, M., & Lombrozo, T. (2017). Ockham's razor cuts to the root: Simplicity in causal explanation. Journal of Experimental Psychology: General, 146, no. 12: 1761-1780.
Powers, L. H. (1978). “Knowledge by Deduction.” Philosophical Review 87, no. 3: 337–371.
Schwartz, D. L., and Black, T. (1999). “Inferences Through Imagined Actions: Knowing by Simulated Doing.” Journal of Experimental Psychology: Learning, Memory, and Cognition 25, no. 1: 116.
Shepard, R. N. (2008). “The Step to Rationality: The Efficacy of Thought Experiments in Science, Ethics, and Free Will.” Cognitive Science 32: 3–35.
Sloman, S. A. (1996). “The Empirical Case for Two Systems of Reasoning.” Psychological Bulletin 119, no. 1: 3–22.

Sobel, D. M., Yoachim, C. M., Gopnik, A., Meltzoff, A. N., and Blumenthal, E. J. (2007). “The Blicket Within: Preschoolers’ Inferences About Insides and Causes.” Journal of Cognitive Development 8, no. 2: 159–182.

Stuart, M. T. (2016). Norton and the logic of thought experiments. Axiomathes, 26, no. 4:, 451-466.

Toon, A. (2010). “Models as Make-Believe.” Beyond Mimesis and Convention: Representation in Art and Science, edited by R. Frigg and M. C. Hunter, 71–96. Dordrecht: Springer.

Walker, C. M., Bonawitz, E., & Lombrozo, T. (2017). Effects of explaining on children's preference for simpler hypotheses. Psychonomic Bulletin & Review, 24, no. 5, 1538-1547.

Walker, C. M., & Lombrozo, T. (2017). Explaining the moral of the story. Cognition, 167: 266-281.

Walker, C. M., Lombrozo, T., Legare, C., and Gopnik, A. (2014). “Explaining Prompts Children to Privilege Inductively Rich Properties.” Cognition 133: 343–357.

Walker, C. M., Lombrozo, T., Williams, J. J., Rafferty, A. N., & Gopnik, A. (2017). Explaining constrains causal learning in childhood. Child Development , 88, no. 1: 229-246.

Weisberg, M. (2012). Simulation and Similarity: Using Models to Understand the World. Oxford: Oxford University Press.

Wilkenfeld, D. A., & Lombrozo, T. (2015). Inference to the best explanation (IBE) versus explaining for the best inference (EBI). Science & Education , 24, no. 9-10: 1059–1077 .

Williams, J. J., and Lombrozo, T. (2010). “The Role of Explanation in Discovery and Generalization: Evidence from Category Learning.” Cognitive Science 34: 776–806.

Williams, J. J., and Lombrozo, T. (2013). “Explanation and Prior Knowledge Interact to Guide Learning.” Cognitive Psychology 66: 55–84.

Williams, J. J., Lombrozo, T., and Rehder, B. (2013). “The Hazards of Explanation: Overgeneralization in the Face of Exceptions.” Journal of Experimental Psychology 142: 1006–1014.