CellPress

Review

# Explanatory Preferences Shape Learning and Inference

Tania Lombrozo[1,*]

**Explanations play an important role in learning and inference. People often learn by seeking explanations, and they assess the viability of hypotheses by considering how well they explain the data. An emerging body of work reveals that both children and adults have strong and systematic intuitions about what constitutes a good explanation, and that these explanatory preferences have a systematic impact on explanation-based processes. In particular, people favor explanations that are simple and broad, with the consequence that engaging in explanation can shape learning and inference by leading people to seek patterns and favor hypotheses that support broad and simple explanations. Given the prevalence of explanation in everyday cognition, understanding explanation is therefore crucial to understanding learning and inference.**

## What Makes a Good Explanation?

In 2012, the Edge Foundation posed the following question to dozens of academics, writers, artists, and intellectuals: 'What is your favorite deep, elegant, or beautiful explanation?' They received nearly 200 responses on a variety of topics, ranging from evolutionary biology and theoretical physics to economics and psychology. Despite this diversity in topics, there was surprising consistency in the reasons respondents offered for favoring their chosen explanations. Appeals to simplicity were ubiquitous, and some respondents appealed to generality or breadth. 'The hallmark of a deep explanation', wrote physicist Max Tegmark, 'is that it answers more than you ask' [1].

The idea that good explanations are simple and broad is not new; it is voiced repeatedly in discussions of explanation from philosophy and the history of science, and it is often espoused by scientists themselves. The philosopher Herbert Feigl, for example, wrote that scientific explanation aims for 'the comprehending of a maximum of facts and regularities in terms of a minimum of theoretical concepts and assumptions' [2]. Albert Einstein wrote that 'the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience' [3].

Strong opinions about what constitutes a good explanation are not limited to intellectuals, philosophers, and scientists. Over the past few decades, work in cognitive, developmental, and social psychology has revealed that untutored children and adults have such opinions as well [4–10]. Moreover, they similarly value simplicity and breadth as 'explanatory virtues' or properties that raise the quality of explanations. How can these virtues be characterized more precisely? And how (if at all) do they influence cognitive processes that often involve explanation, such as inference and learning?

One reason it is important to answer these questions is because explanation plays an important role in many everyday cognitive processes [11–16]. For instance, we might infer that the butler 'did it' because this hypothesis best explains the data, or come to understand an opponent's

### Trends

Children and adults have systematic preferences for some explanations over others. They favor explanations with 'explanatory virtues', such as simplicity and breadth, especially when direct probabilistic evidence is unavailable.

Explanatory virtues are difficult to quantify and precisely define, but advances in philosophy and experimental psychology are reinvigorating the study of explanations and their role in cognition.

Explanatory preferences are consequential because explanation plays a crucial role in inference and learning. We often infer to the best explanation, or learn by explaining to others or ourselves. When people engage in such explanation-based processes, explanatory virtues are recruited as evaluative constraints. As a result, explaining can facilitate the discovery of simple hypotheses and broad patterns, but it can also lead to systematic errors.

[1]Department of Psychology, University of California, Berkeley, CA, USA

*Correspondence: lombrozo@berkeley.edu (T. Lombrozo).

 http://dx.doi.org/10.1016/j.tics.2016.08.001

chess move by trying to explain why it was chosen. If explanatory preferences influence how these processes unfold, then the study of explanation becomes crucial to the study of learning and inference.

In the past few years research in psychology and philosophy has begun to address explanation and its role in cognition in new and exciting ways. Synthesizing this emerging body of work, I and my colleagues have suggested that when children and adults generate and evaluate explanations, they recruit explanatory virtues, such as simplicity and breadth, as evaluative constraints on reasoning [5] (see also [17–19]). As a result, they are more likely to generate and favor broad and simple hypotheses, and to discover broad and simple patterns, for better or for worse [20].

This article reviews major recent developments in the study of the explanatory preferences of children and adults, with particular attention to simplicity and breadth. Each virtue supports two complementary projects, reviewed in turn. The first is to characterize a given virtue more precisely and to determine whether that virtue–so defined–in fact influences judgments of explanation quality. The second project is to investigate the downstream consequences of invoking that virtue when engaged in explanation-based processes, such as inference to the best explanation (Box 1) or learning through self-explanation (Box 2). Although the second project is conceptually dependent upon the first, the two can be mutually constraining [5]: if engaging in explanation recruits explanatory virtues as constraints on processing, then one can expect the process of explaining to magnify the influence of explanatory virtues on learning and inference. The effects of explanation thus provide clues about the nature of explanatory preferences, and documented preferences provide a basis for predicting the downstream effects of explanation.

### Box 1. Inference to the Best Explanation

In a 1965 paper, the philosopher Gilbert Harman coined the term 'inference to the best explanation' (IBE), which he characterized as an inference "from the premise that a given hypothesis would provide a 'better' explanation for the evidence than would any other hypothesis, to the conclusion that the given hypothesis is true" [48]. IBE can be generalized from an inference concerning a single conclusion to a more general rule for assigning subjective probabilities to explanatory hypotheses, where an explanation's 'loveliness'–for instance, its simplicity and breadth–guides judgments about its 'likeliness' [49].

One challenge for advocates of IBE has been to more precisely characterize explanatory virtues: the features that make for 'loveliness'. Another challenge has been to articulate the relationship between IBE and Bayesian inference, and, if they diverge, to justify why IBE could ever be warranted (e.g., [50]). Against this backdrop, two developments in philosophy of science and epistemology are worth highlighting.

First, work in philosophy of science demonstrates how the types of explanatory considerations that motivate IBE could be compatible with Bayesian inference, and in fact emerge from Bayesian inference over appropriate structures [51,52]. One example of this idea can be seen in 'Bayesian Occam's razor', roughly the idea that simpler hypotheses–in this case meaning those that are less flexible–will receive a probabilistic boost because the prior probability assigned to the corresponding set of hypotheses will not be distributed over as many possibilities [53]. Within this approach, IBE is not only compatible with Bayesian inference, but in fact responsive to the same epistemic considerations [51].

Second, some contemporary formal epistemologists defend 'explanationism' as a probabilistic alternative to Bayesianism [54], and identify conditions under which explanatory judgments can lead to inferences that are more accurate than those based on the application of Bayes' rule [55,56]. To understand this result, it helps to remember the sense in which Bayes' rule is an optimal inference rule: it minimizes average, long-term inaccuracy. With a different epistemic goal–such as being mostly right in the short term–an alternative rule for updating beliefs can outperform Bayes' rule, as shown in a series of simulations [55,57]. Moreover, models that incorporate estimates of explanation quality offer a better descriptive characterization of people's inferences than do models that compute posterior probability directly ([58,59], see also [60]).

These developments support an important role for IBE in describing people's everyday cognition and in achieving important epistemic goals. However, they differ in whether they assimilate IBE to Bayesian inference or instead offer it as an alternative. It may be that each approach is partially correct, depending on the explanatory virtue in question.

### Box 2. Learning by Explaining

Explanation and learning are intimately related. In educational contexts, explanations can transmit information from instructor to student (e.g., [61]), and are used to assess student understanding (e.g., [62]). Moreover, the very process of explaining, even without feedback, can have powerful effects. This phenomenon is known as the self-explanation effect [63,64] and has been documented in a variety of contexts [65]. For instance, one study found that prompting undergraduate students to explain worked examples of fraction division out loud to themselves improved conceptual learning more effectively than control prompts [66].

Learning by self-explaining is intriguing as an instance of 'learning by thinking' [67]: genuine insight or learning can occur in the absence of novel observations or evidence. How is this possible? There are several mutually consistent proposals. For instance, explaining could encourage learners to draw inferences to fill gaps in their knowledge, and to revise mental models to better accord with what they are explaining [68,69]. Explanation involves the coordination of what is being explained with prior beliefs (e.g., [19,70]), and can lead to better metacognitive calibration (e.g., [71], see also [72,73]). In addition, as reviewed here, explanation can recruit explanatory virtues as constraints on reasoning, influencing the hypotheses a learner favors and entertains [5]. These processes could help learners to generate accurate explanations, and more generally revise their representation of the problem or domain.

Interestingly, there is evidence that benefits of engaging in explanation are not always a consequence of having thereby acquired a correct explanation. For example, one study prompted 8th-grade students to either explain to themselves as they studied a text about the human circulatory system, or to read the study materials twice [64]. Those who self-explained outperformed the control group on a subsequent test, even though the self-explanations were often incorrect. Michelene Chi and her colleagues suggest that generating an explanation can 'objectify' erroneous assumptions, helping learners to recognize errors and repair their mental models [64]. In addition, there is evidence that explaining incites processes such as comparison [74] and abstraction [18,33], which can improve a learner's representation of a problem or domain even if the processes fall short of producing an accurate explanation. These powerful consequences of explaining motivate the proposal that people engage in a process of 'explaining for the best inference' (EBI), which differs from inference to the best explanation (IBE, Box 1) in focusing on the consequences of explanation as a process rather than a product [75].

## Simplicity

In the *Principia Mathematica*, Newton advises that we 'admit no more causes of natural things than such as are true and sufficient to explain their appearances' (quoted in [21]). Other advocates for simplicity include William of Occam, with his famous 'razor' to trim away complexity, as well as Aristotle, Aquinas, Kant, and many more. Simplicity also finds more contemporary advocates, especially in the context of model selection (e.g., [22]). However, not all advocates operate with the same notion of simplicity. Indeed, simplicity is anything but simple to define. An explanation could be simple in the sense that it appeals to few actual entities or to few different types of entities, in the sense that it supports a concise description, or in the sense that it is highly inflexible, to list only a few possibilities [21,23].

Within psychology, only a handful of studies have directly investigated how people choose between competing explanations that differ with respect to some measure of simplicity. Early studies [24] tested a measure endorsed by the philosopher Paul Thagard, according to which simpler theories (and by extension simpler explanatory propositions) 'make fewer special assumptions' [25]. Participants were thus asked to select between explanations that accounted for a set of observations equally well, but differed in the number of assumptions they required. For instance, Cheryl's recent nausea, weight gain, and fatigue could be explained with the single assumption that she is pregnant, or instead with the conjunction of three independent assumptions: that she has a stomach virus, has stopped exercising, and has mononucleosis. Participants tended to favor one of the assumptions in the conjunctive explanation when presented with a single symptom (e.g., they preferred to explain Cheryl's nausea by appeal to a stomach virus when nausea was her only symptom), but when all three symptoms were present, they strongly favored the 'simpler' explanation (pregnancy).

These findings provide *prima facie* support for the idea that people favor simpler explanations, where simplicity is a matter of minimizing independent assumptions. However, they could also be a consequence of straightforward probabilistic considerations. For example, if the candidate

causes are all rare and have comparable probabilities, it is more likely that Cheryl would have only one rather than a conjunction of three. Subsequent work has therefore provided participants with probabilistic evidence, allowing the experimenter to effectively control the relative probabilities of candidate explanations. This research reveals that when the probability of each candidate explanation is explicitly stipulated, participants favor the more likely explanation, even when it involves two causes instead of one [26]. However, when probabilistic cues are indirect and some uncertainty remains, explanation choices are a function of both simplicity and probability.

In one set of studies [26], participants selected between explanations for an individual's symptoms that appealed to either a single common cause (simple) or to two independent causes (complex), and they were additionally provided with information about the base rate of each disease. Under these conditions, explanation choices were sensitive to both simplicity and probability. When two candidate explanations were equally likely, a majority of participants selected the simpler explanation. It was not until the complex explanation was 10-fold more likely than the simpler alternative that a majority of participants selected it as the most satisfying explanation for the symptoms. It thus appears that, when the probability of an explanation is uncertain, the relative simplicity of an explanation has a significant effect on its perceived quality.

Even young children use simplicity as a basis for choosing between competing explanations. In a study with children aged 4–6 years [27], the children mirrored the adults from [26] in favoring a common-cause explanation over an explanation invoking two independent causes, and in responding as a function of both simplicity and probability. Interestingly**,** adults did not show any preference for simpler explanations in this child-friendly task, which involved more transparent numbers and causal mechanisms than prior work [26]. This supports the idea that simplicity informs inference only when more transparent and reliable guides to probability are absent.

More recent work[i] (cited in [16]) has sought to clarify what it is that makes a common cause explanation 'simpler' than an alternative that invokes multiple independent causes. One possibility is that participants compare the number of causes invoked in an explanation. Another possibility is that participants favor explanations that involve the fewest assumptions, in this case the fewest causes that are themselves unexplained. Experiments explicitly contrasting these two measures of simplicity find strong support for the latter possibility, and no support for the former: participants favored explanations that invoked the fewest causes that were themselves unexplained, even when that explanation invoked more causes overall (Figure 1). As in [26], this preference traded off with information about probability.

The cases considered so far concern causal inference: general causal relationships are already known (e.g., that pregnancy can cause nausea), and participants are tasked with identifying the best explanation for an individual case (e.g., Cheryl's present nausea). However, there is also evidence that a preference for simplicity could play a role in causal learning, where participants learn which causal relationships exist by observing multiple cases. Specifically, there is evidence that when inferring which of multiple candidate causes can generate an effect, participants approach the causal learning problem with 'generic priors' that favor a small number of strong causes ([28,29],but see also [30]).

In sum, recent experimental work suggests that people favor simpler explanations, and that this preference manifests in the context of both causal inference and causal learning. Although it is likely that multiple metrics for simplicity operate in parallel, the evidence to date supports two proposals: that explanations for individual events are better to the extent they invoke fewer independent assumptions, and that when learning multi-causal systems, learners favor sparse and strong causal structures.

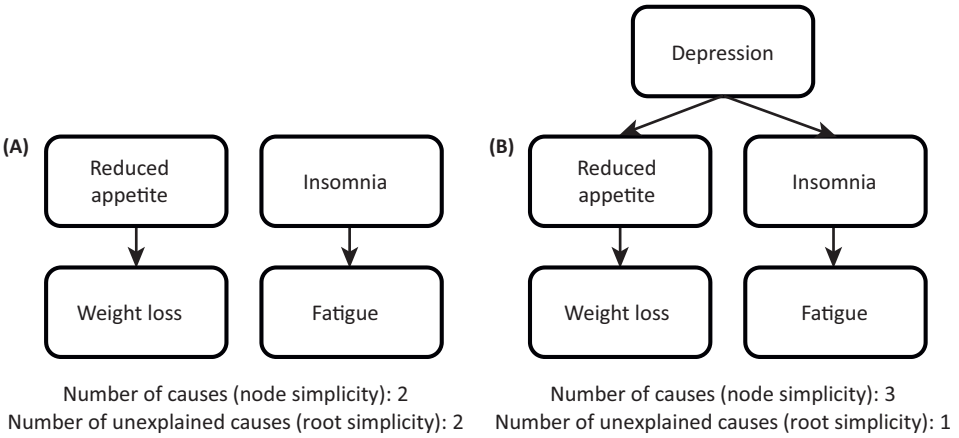


Figure 1. Differentiating Measures of Simplicity in Causal Explanation. Consider a patient suffering from weight loss and from fatigue. One explanation (A) is that these symptoms result from reduced appetite and insomnia, where these causes are themselves unexplained (and in particular, not caused by the common cause of depression). A second explanation (B) is that depression is responsible for the reduced appetite and insomnia, which in turn caused the symptoms. In a study[i] reviewed in [16], participants were presented with fictional scenarios employing such structures which were designed to differentiate the total number of causes invoked in an explanation ('node simplicity') from the number of causes that were themselves unexplained ('root simplicity'). In this case, explanation (A) has lower node simplicity, but explanation (B) has lower root simplicity. Before making an explanation choice about a particular individual, participants saw many sample diagnoses, which effectively communicated the joint probability distribution over causes. If participants were insensitive to both node and root simplicity, explanation choices should have been a function of this probabilistic information–but instead responses were also influenced by root simplicity. When the root-simpler explanation was less probable than the alternative, participants selected the root-simpler explanation significantly more often than warranted by the probability information alone. There was no evidence that explanations with lower node simplicity (that is, that invoked a smaller number of causes) were favored.

## Consequences for Learning and Inference

If children and adults favor explanations that are simpler in the sense that they invoke fewer assumptions or unexplained causes, then actively engaging in explanation-seeking could magnify this preference in a causal inference task. To test this prediction, children aged 4–6 years were presented with two observations that could be explained by appeal to a single common cause or by appeal to two independent causes [31]. Half the children were prompted to explain the two observations, and half to report what they were. A later inference task revealed that 5-year-old children who were prompted to explain the observations were significantly more likely to make inferences in line with the simpler, common-cause hypothesis than were 5-year-olds prompted to report. By contrast, 4-year-olds responded at chance in both conditions, and 6-year-olds favored the simpler hypothesis whether or not they were prompted to explain.

These findings reveal that engaging in explanation can exaggerate the influence of simplicity on reasoning. However, they also suggest some boundary conditions: the predicted effect was only observed for participants in the middle of the age-range tested. It is likely the younger children found the task too demanding, but the shift between age 5 and 6 years is less clear. One possibility is that older children engaged in explanation spontaneously (see [32]), attenuating the effects of the experimental manipulation. Another possibility is that explaining helped 5-year-olds overcome the appeal of more salient and familiar causes, unmasking their explanatory preference, whereas 6-year-olds were better able to resist the salient and familiar on their own (see [33]).

More subtle measures have revealed a complementary effect of explanation in adults [26]. Participants were required to choose the most satisfying explanation for an individual's symptoms, and also to recall data they had observed earlier in the task about the frequencies with

which candidate diseases (co)occurred. Participants who chose a simple explanation when it was unlikely to be true also tended to systematically over-report the prevalence of the disease invoked in that explanation. However, this effect was greater when the explanation task came before the memory task, suggesting a causal influence of engaging in explanation on subsequent memory[i] [16].

The studies reviewed above used well-controlled experiments to document systematic effects of engaging in explanation on the role of simplicity in reasoning, but it is likely that, in more-realistic settings, such effects are even more pronounced. People may be more likely to spontaneously entertain simple hypotheses (failing to recognize complex alternatives when they are not made salient), and to more readily remember and communicate them. For those interested in improving real-world decision making, understanding the effects of a preference for simpler explanations may be all the more important.

## Boundary Conditions

Simpler explanations may not be judged better along all dimensions. In particular, there is evidence that complexity may be better when it comes to evaluating plausibility or establishing expertise, although work demonstrating these effects has quantified simplicity in variable ways, and has not always measured explanatory judgments directly. For example, arguments with more premises (and hence more assumptions) are sometimes taken to better support their conclusions [34] and, in one study, mock jurors were more influenced by an article when the article used complex as opposed to simple language [35]. There is also evidence that in the context of curve fitting, more-complex (i.e., higher-order) lines are sometimes favored because their complexity creates the sense that they provide a better fit to the data [36]. Finally, a set of studies that measured explanation judgments directly found that adults preferred explanations that were longer [37] (Box 3). These results reinforce the point that 'simplicity' is likely to

### Box 3. Explanatory Vices

Explanatory virtues are features of explanations that increase their perceived quality, but they often have the additional normative implication that this influence is appropriate. Explanatory vices, then, can be defined as features of explanations that inappropriately influence their perceived quality (see also [76]).

One example is the 'seductive allure' effect [77], first demonstrated in the context of psychological explanations. Both experts and non-experts evaluated good and bad explanations for psychological phenomena, where the explanations also varied in whether they included some mostly-irrelevant neuroscientific jargon. Although participants were good at discriminating the good from the bad explanations overall, non-experts did so less effectively when the neuroscientific information was added. In particular, the irrelevant neuroscience made the bad explanations seem less bad. Subsequent work has revealed that both the length of an explanation and references to the brain contribute to the perceived quality of psychological explanations [37]. Moreover, the effect is unlikely to be unique to psychological explanations. Other 'reductive' scientific explanations have the same effect [78], and adding irrelevant math to a scientific abstract improves the perceived quality of the research [79].

A second example comes from work on acceptance by children and adults of scientifically questionable teleological explanations. For instance, young children will happily explain that mountains are for climbing, and that clouds are for producing rain [80]. Under speeded conditions, adults similarly err in the direction of accepting unwarranted teleological explanations, such as 'the sun makes light so that plants can photosynthesize' [81]. Although adults with scientific training are less likely to accept such explanations overall, they similarly err towards acceptance when responding under speeded conditions [82].

These explanatory vices may be side-effects of reasonable heuristics for assessing explanation quality. In many cases, scientific jargon is a sign of expertise. Thus laypeople may defer to the explanation-provider, and assume that the jargon contains or points to information that would provide an adequate explanation. Similarly, longer explanations are typically more informative. The case for teleological explanations is less clear. Some argue that teleological explanations reflect an 'explanatory default' that is suppressed, but not replaced, by non-teleological, scientific alternatives ([81,82], see also [83] for relevant discussion). Another possibility is that the preference reflects the operation of a defeasible heuristic, whereby a good 'fit' between a structure and a function provides defeasible evidence that a teleological explanation is warranted [84].

correspond to more than one feature of explanations, and that it will certainly trade off against other features in particular cases.

## Breadth

Intuitively, better explanations are broader explanations: they explain a broader range of observations or phenomena. This idea is captured by a variety of proposals within psychology and philosophy, and under a variety of different names, including scope [18] and coverage [38]. However, as with simplicity, making the notion of breadth more precise is not straightforward. For instance, is an explanation broader if it predicts a larger number of observations than an alternative, but with lower probability or with less precision? What if an explanation accounts for fewer but more diverse observations? Most research has focused on cases where such tradeoffs do not arise, with the consequence that breadth has remained underspecified.

Breadth finds one articulation by Paul Thagard: 'Other things being equal, we should prefer a hypothesis that explains more than alternative hypotheses. If hypothesis H1 explains two pieces of evidence while H2 explains only one, then H1 should be preferred to H2' [39]. Working with this definition, one study [24] presented participants with two or three facts that could be explained by appeal to a 'narrow' explanation (which explained only one) or a 'broad explanation' (which explained them all). For instance, to explain Cheryl's nausea, weight gain, and fatigue, participants would evaluate 'she has a stomach virus' (which explains only nausea) and 'she is pregnant' (which explains all three symptoms). Although participants favored narrow explanations when presented with the corresponding single facts, the broad explanations were strongly favored when multiple facts were presented.

As with simplicity, it can be difficult in such cases to disentangle a preference for breadth as such from more direct probabilistic considerations. In a case such as Cheryl's, it is very plausible that the three symptoms provide stronger evidence for the broad explanation than for the narrow explanation. With some additional assumptions about the base rates of the various conditions, this renders the preference for the broader explanation a straightforward consequence of probabilistic inference. Similarly, there is evidence that people favor diagnostic inferences that explain a diverse set of symptoms over those that explain an equal number of more closely related symptoms [40], but such effects of diversity can potentially be explained as a consequence of normative, probabilistic inference over causal models [41]. Findings with children are similarly ambiguous (e.g., [42]).

Indirect evidence, however, does suggest that explanatory considerations related to breadth can diverge from those related to probability. In one set of studies, people found an explanatory belief more valuable when prompted to consider all the observations that the belief could explain, without comparable effects on perceived probability [43]. The measure of value used in these studies included how important, meaningful, personally relevant, and societally impactful the explanatory belief seemed to be, but did not include a direct measure of its perceived quality as an explanation. Another study [44] found that explanations (diseases) that accounted for three observed symptoms were judged better explanations than those that accounted for only one, even when the presence of the disease was stipulated. This suggests that, if the relevant consideration is the amount of evidence that the observations provide for the explanation, then the value of evidence extends beyond its role in supporting an inference to the truth of the explanation.

In sum, there is evidence that adults favor explanations that account for more evidence over those that account for less, but it remains unclear whether this preference is a direct consequence of probabilistic inference, or if it instead reflects a preference for breadth as such. As we will see below, the strongest evidence might come from experiments investigating the effects of

engaging in explanation, which suggest that explaining magnifies children's and adults' preference for breadth, with significant consequences for learning and inference.

## Consequences for Learning and Inference

If people favor broader explanations, then actively seeking an explanation could bias learners towards hypotheses that support broad explanations. Recent work supports this prediction. In one set of studies [18], participants were tasked with learning two novel categories by studying four exemplars from each (Figure 2). Participants were either prompted to explain why each study item might belong to its respective category (without receiving feedback), or, in a control task, to describe the item, think aloud while studying it, or engage in free study. Participants who were prompted to explain were significantly more likely than those in any control group to discover a subtle rule that accounted for the membership of all eight study items. Those in the control conditions tended to notice a salient but imperfect rule that accounted for six of the eight items, without successfully discovering the broader alternative.

Subsequent work has extended this finding to richer contexts [19] and to causal learning in children [17], and has also identified conditions under which explaining can hinder learning, rather than help [20]. For instance, in one study participants were prompted to either explain the category membership of study items or to engage in a control task [20]. This time, however, half the participants were learning novel categories that supported a broad rule (i.e., one that accounted for all observations), and half were presented with 'misleading' regularities, such

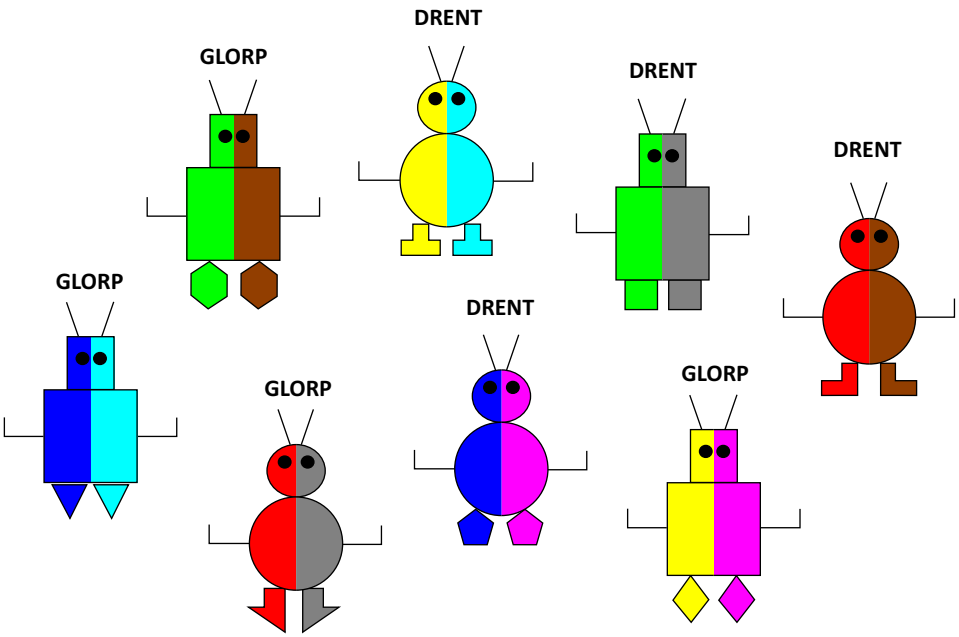


Figure 2. Stimuli from a Categorization Task Illustrating the Effects of Explanation on Learning. In one set of studies [18], participants were presented with eight exemplars similar to those presented here, four of which belonged to the category 'glorp' and four to the category 'drent'. The exemplars were designed to support two potential classification rules. The first, more salient rule concerned body shape, with the majority of glorps having square bodies, and the majority of drents having round bodies. However, this '75% rule' only captured 75% of the exemplars. The second, more subtle '100% rule' concerned foot shape: all glorps had feet that were pointy on the bottom, and all drents had feet that were flat along the bottom. As participants studied these exemplars, they were either prompted to explain why each item might belong to its respective category or to engage in a control task: description, thinking out loud, or free study. Across three experiments, those participants who were prompted to explain were significantly more likely than those in the control group to discover the 100% rule (a version of this figure appeared originally in [18]).

that the only way to achieve perfect classification was to memorize idiosyncratic properties of individual exemplars. In the former case, participants prompted to explain tended to learn to categorize more quickly and with fewer errors than those in the control conditions. By contrast, in the latter case, explainers were significantly slower and made significantly more errors. Explainers seemed to perseverate in looking for a good explanation–a simple, broad rule–even when no such rule was available.

Documenting cases in which engaging in explanation can hinder learning is especially valuable for two reasons. First, as with visual illusions, such cases can be highly diagnostic of underlying mechanisms. Cases in which explanation improves learning could potentially be reduced to effects of attention or engagement, but the finding that explanation leads learners to perseverate in seeking broad patterns–even when doing so hinders learning–provides unique support for the idea that explaining recruits the virtue of breadth. Second, identifying such cases has practical value: self-explanation prompts are often used in pedagogical contexts (Box 2), and this work helps to specify when and why they are likely to be effective.

## Boundary Conditions

In many real-world situations, explanations are evaluated under conditions of uncertainty, and uncertainty appears to change the value of broad scope. For instance, suppose you are a doctor trying to diagnose a patient who could have either Vellereum or Pythium. Vellereum always leads

### Box 4. Beyond Simplicity and Breadth

Within philosophy, a variety of explanatory virtues have been proposed, including a specification of mechanism, precision, fruitfulness, and fit with background beliefs [49]. However, few of these have been investigated empirically. Two exceptions are below.

*Explanatory Power*

One recent development is a Bayesian analysis of 'explanatory power', intended to capture what renders one candidate explanation stronger than another [85]. The analysis begins by identifying intuitive adequacy conditions–for instance, that all else being equal, an explanation has greater power than an alternative if it makes what is being explained less surprising. A single measure uniquely captures these conditions, where the power of an explanatory hypothesis $h$ with respect to a target of explanation $t$ can be expressed as:

$$\frac{p(h|t)-p(h|\neg t)}{p(h|t)+p(h|\neg t)}$$

Subsequent work has shown that this measure captures people's intuitive judgments of explanation quality in a simple probabilistic scenario more effectively than alternatives [59].

*Mechanisms*

Recent proposals within philosophy tie explanations to mechanisms [86], and within psychology there is also suggestive evidence. For instance, people tend to seek information about mechanisms when engaged in causal attribution [87]. Moreover, explanation statements ('Why Q? Because P') receive higher ratings when participants are presented with a possible mechanism linking P to Q, and this mechanistic information has a greater impact on ratings of explanation claims than on closely matched causal claims about P causing Q [88].

Additional evidence comes from development. In one task [32], children aged 3–6 years were more likely to learn the mechanism by which a novel toy made a fan spin when they explained how it worked, whether the explanation was prompted or spontaneous. This learning benefit was selective: explaining improved memory for mechanism, but not for causally irrelevant details. In another task [33], children aged 3–5 years observed three blocks placed one at a time on a machine that either did or did not activate. Half the children were prompted to explain why each block did (or did not) activate the machine; the other half to report whether it did (or did not). The experimenter then revealed that one block that had activated the machine had an internal part. The experimenter asked the child to identify which of the other blocks had the same part inside, where one was perceptually identical but had not activated the machine, and the other was perceptually distinct but had. Children who explained were significantly more likely to generalize the internal part to the block with the same causal property, suggesting that generating explanations had prepared kids to assimilate and extend information about unobserved mechanisms.

to abnormal levels of alanine in the blood; Pythium always leads to abnormal levels of both alanine and valine. You know your patient has abnormal levels of alanine, but the test results for valine have not yet come in. Which is more likely, that your patient has Vellerium or Pythium?

This question forces participants to choose between candidate explanations that differ in their 'latent' scope–that is, in the number of unverified predictions that they make. Across several studies using stimuli like these [45], participants exhibited a reliable preference for explanations with narrower latent scope (e.g., favoring Vellerium over Pythium), an effect that has been subsequently replicated with children [46]. One proposal, for which there is some support [47], is that people make an inference about the unverified prediction (e.g., that the patient's valine levels are probably not abnormal because abnormal levels are rare), and then use the inferred evidence as a basis for explanation choice. On this view, it is not broad latent scope *per se* that penalizes explanations, but instead making unverified predictions that are abnormal or unlikely.

## Concluding Remarks and Future Directions

The past decade of research in cognitive and developmental psychology has made important advances in the study of explanation, going beyond the intuitive platitude that simpler and broader explanations are better explanations to a more systematic investigation of what simplicity and breadth amount to, and why these explanatory preferences matter. These empirical advances have been accompanied by advances in formal epistemology and philosophy of science that help to characterize explanatory virtues more precisely, including the relationship between explanation-based inference and inferences based on the application of Bayes' rule (Boxes 1 and 4).

Although much work remains to be done (see Outstanding Questions), current research makes a strong case for the value of explaining explanation in understanding cognition. Explanation plays a central role in many everyday cases of inference and learning, and as a result explanatory preferences exert substantive constraints on cognitive processes, shaping the types of structure people seek and the hypotheses they favor. Explaining explanation is thus an invaluable step towards a more comprehensive understanding of our remarkable capacity to reason and learn.

### Acknowledgments

The author wishes to acknowledge Elizabeth Kon and Jonah Schupbach for helpful comments and discussion, and several sources for support: National Science Foundation (NSF) CAREER grant DRL-1056712, a James S. McDonnell Foundation Scholar Award in Understanding Human Cognition, and the Templeton Foundation's Varieties of Understanding project.

### Resources

[i] M. Pacer and T. Lombrozo, Ockham's razor cuts to the root: simplicity in causal explanation (July 29, 2016). Available at SSRN: http://ssrn.com/abstract = 2815758.

### References

1. Brockman, J. (2013) *This Explains Everything,* Harper Perennial
2. Feigl, H. (1970) The 'orthodox' view of theories: remarks in defense as well as critique. In *Minnesota Studies in the Philosophy of Science* (4) Radner, M. and Winokur, S.,eds In pp. 3–16, University of Minnesota Press
3. Einstein, A. (1934) On the method of theoretical physics. *Philos. Sci.* 1, 163–169
4. Lombrozo, T. (2011) The instrumental value of explanations. *Philos. Compass* 6, 539–551
5. Lombrozo, T. (2012) Explanation and abductive inference. In *The Oxford Handbook of Thinking and Reasoning* (Holyoak, K.J. and Morrison, R.G., eds), pp. 260–276, Oxford University Press
6. Lombrozo, T. (2016) Explanation. In *A Companion to Experimental Philosophy* (Systma, J. and Buckwalter, W., eds), pp. 491–503, John Wiley & Sons
7. Wellman, H.M. (2011) Reinvigorating explanations for the study of early cognitive development. *Child Dev. Perspect.* 5, 33–38
8. Frazier, B.N. *et al.* (2009) Preschoolers' search for explanatory information within adult-child conversation. *Child Dev.* 80, 1592–1611
9. Frazier, B.N. *et al.* (2016) Young children prefer and remember satisfying explanations. *J. Cogn. Dev.* Published online February 23, 2016. http://dx.doi.org/10.1080/15248372.2015.1098649
10. Keil, F.C. (2011) Explanation and understanding. *Annu. Rev. Psychol.* 57, 227–254
11. Hastie, R. and Pennington, N. (1999) Explanation-based decision making. In *Judgment and Decision Making: An Interdisciplinary Reader* (Connolly, T. *et al.*, eds), pp. 212–228, Cambridge University Press
12. Lombrozo, T. (2009) Explanation and categorization: how 'why?' informs 'what?'. *Cognition* 110, 248–253

### Outstanding Questions

People often favor explanations that are simple and broad, but to what extent do these preferences vary across domains and individuals? For example, do people have stronger preferences for simplicity in some domains than in others? How do individual differences in explanatory preferences relate to other factors, such as culture and expertise?

Much of the evidence for explanatory preferences comes from artificial cases in which explanations differ along a single dimension. However, in real-world cases virtues often compete: the simplest explanation need not be the broadest. How do multiple virtues trade off to determine global explanation preferences, and how do these virtues interact with probabilistic evidence?

When and why is it rational to use explanatory considerations as a guide to learning and inference? In some cases, favoring 'virtuous' explanatory hypotheses can improve learning and inference, whereas in other cases the effects are detrimental. Descriptive theories of the role of explanation in cognition would benefit from normative counterparts that provide a benchmark for optimal performance, that allow us to better appreciate the relationship between explanation-based processes and other forms of inference (such as Bayesian inference), and that help to differentiate explanatory virtues that serve a cognitive function from those that are side-effects of other mechanisms or representations.

Prototypical explanations are explicit, verbal responses to some form of query. Nonetheless, explanations can depart from this prototype, and other cognitive products and processes share elements with explanation as well. To what extent do 'explanatory' preferences shape forms of learning and inference that are more implicit?

Where do explanatory preferences come from? This question can be asked at multiple levels: in terms of biological evolution, cultural evolution, and human development. If explanatory preferences are learned, what are the mechanisms involved?

13. Lombrozo, T. and Gwynne, N.Z. (2014) Explanation and inference: mechanistic and functional explanations guide property generalization. *Front. Hum. Neurosci.* 8, 700
14. Legare, C.H. (2014) The contributions of explanation and exploration to children's scientific reasoning. *Child Dev. Perspect.* 8, 101–106
15. Khemlani, S. and Johnson-Laird, P.N. (2013) Cognitive changes from explanations. *J. Cogn. Psychol.* 25, 139–146
16. Lombrozo, T. and Vasilyeva, N. Causal Explanationexplanation. In *Oxford Handbook of Causal Reasoning* (Waldmann, M., ed.), Oxford University Press (in press)
17. Walker, C.M. *et al.* (2016) Explaining constrains causal learning in childhood. *Child Dev.* Published online July 8, 1016. http://dx.doi.org/10.1111/cdev.12590
18. Williams, J.J. and Lombrozo, T. (2010) The role of explanation in discovery and generalization: evidence from category learning. *Cogn. Sci.* 34, 776–806
19. Williams, J.J. and Lombrozo, T. (2013) Explanation and prior knowledge interact to guide learning. *Cogn. Psychol.* 66, 55–84
20. Williams, J.J. *et al.* (2013) The hazards of explanation: overgeneralization in the face of exceptions. *J. Exp. Psychol. Gen.* 142, 1006–1014
21. Baker, A. (2013) Simplicity. In *The Stanford Encyclopedia of Philosophy* (Zalta, E.N., ed.),
22. Forster, M. and Sober, E. (1994) How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *Br. J. Philos. Sci.* 45, 1–35
23. Sober, E. (2015) *Ockham's Razors: A User's Manual,* Cambridge University Press
24. Read, S.J. and Marcus-Newhall, A. (1993) Explanatory coherence in social explanations: a parallel distributed processing account. *J. Pers. Soc. Psychol.* 65, 429–447
25. Thagard, P. (1989) Explanatory coherence. *Behav. Brain Sci.* 12, 435–467
26. Lombrozo, T. (2007) Simplicity and probability in causal explanation. *Cogn. Psychol.* 55, 232–257
27. Bonawitz, E.B. and Lombrozo, T. (2012) Occam's rattle: children's use of simplicity and probability to constrain inference. *Dev. Psychol.* 48, 1156–1164
28. Lu, H. *et al.* (2008) Bayesian generic priors for causal learning. *Psychol. Rev.* 115, 955–984
29. Powell, D. *et al.* (2016) Causal competition based on generic priors. *Cogn. Psychol.* 86, 62–86
30. Yeung, S. and Griffiths, T.L. (2015) Identifying expectations about the strength of causal relationships. *Cogn. Psychol.* 76, 1–29
31. Walker, C. *et al*. Effects of explaining on young children's preference for simpler hypotheses. *Psychon. Bull. Rev*. (in press)
32. Legare, C.H. and Lombrozo, T. (2014) Selective effects of explanation on learning during early childhood. *J. Exp. Child Psychol.* 126, 198–212
33. Walker, C.M. *et al.* (2014) Explaining prompts children to privilege inductively rich properties. *Cognition* 133, 343–357
34. Heit, E. and Rotello, C.M. (2012) The pervasive effects of argument length on inductive reasoning. *Think. Reason.* 18, 244–277
35. Lawson, V. (2014) *The Influence of Naive and Media-Informed Beliefs on Juror Evaluations of Forensic Science Evidence,* CUNY
36. Johnson, S.G.B. *et al.* (2014) Simplicity and goodness-of-fit in explanation: the case of intuitive curve-fitting. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (Bello, P. *et al.*, eds), pp. 701–706, Austin, TX, Cognitive Science Society
37. Weisberg, D.S. *et al.* (2015) Deconstructing the seductive allure of neuroscience explanations. *Judgm. Decis. Mak.* 10, 429–441
38. Pennington, N. and Hastie, R. (1992) Explaining the evidence: tests of the story model for juror decision making. *J. Pers. Soc. Psychol.* 62, 189–206
39. Thagard, P. (1992) *Conceptual Revolutions,* Princeton University Press
40. Kim, N.S. and Keil, F.C. (2003) From symptoms to causes: diversity effects in diagnostic reasoning. *Mem. Cognit.* 31, 155–165
41. Rebitschek, F.G. *et al.* (2016) The diversity effect in diagnostic reasoning. *Mem. Cognit* 44, 789–805
42. Samarapungavan, A. (1992) Children's judgments in theory choice tasks: scientific rationality in childhood. *Cognition* 45, 1–32
43. Preston, J. and Epley, N. (2005) Explanations versus applications: the explanatory power of valuable beliefs. *Psychol. Sci.* 16, 826–832
44. Johnson, S.G.B. *et al.* (2014) Explanatory scope informs causal strength inferences. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (Bello, P. *et al.*, eds), pp. 2453–2458, Austin, TX, Cognitive Science Society
45. Khemlani, S.S. *et al.* (2011) Harry Potter and the sorcerer's scope: latent scope biases in explanatory reasoning. *Mem. Cognit.* 39, 527–535
46. Johnston A.M. *et al*. Little Bayesians or little Einsteins?. Probability and explanatory virtue in children's inferences. *Dev. Sci*. (in press)
47. Johnson, S.G.B. *et al.* (2016) Sense-making under ignorance. *Cogn. Psychol.* 89, 39–70
48. Harman, G.H. (1965) The inference to the best explanation. *Philos. Rev.* 74, 88
49. Lipton, P. (2003) *Inference to the Best Explanation,* Routledge
50. Van Fraassen, B.C. (1989) *Laws and Symmetry,* Oxford University Press
51. Henderson, L. (2014) Bayesianism and inference to the best explanation. *Br. J. Philos. Sci.* 65, 687–715
52. Henderson, L. *et al.* (2010) The structure and dynamics of scientific theories: a hierarchical Bayesian perspective. *Philos. Sci.* 77, 172–200
53. MacKay, D.J.C. (2003) *Information Theory, Inference and Learning Algorithms,* Cambridge University Press
54. Douven, I. and Schupbach, J.N. (2015) Probabilistic alternatives to Bayesianism: the case of explanationism. *Front. Psychol.* 6, 459
55. Schupbach, J.N. Inference to the best explanation, cleaned up and made respectable. In *Best Explanations: New Essays on Inference to the Best Explanation* (McCain, K. and Poston, T., eds), Oxford University Press (in press)
56. Douven, I. and Wenmackers, S. (2015) Inference to the best explanation versus Bayes's rule in a social setting. *Br. J. Philos. Sci.* Published online July 31, 2015. http://dx.doi.org/10.1093/bjps/axv025
57. Douven, I. (2013) Inference to the best explanation, Dutch books, and inaccuracy minimisation. *Philos. Q.* 63, 428–444
58. Douven, I. and Schupbach, J.N. (2015) The role of explanatory considerations in updating. *Cognition* 142, 299–311
59. Schupbach, J.N. (2011) Comparing probabilistic measures of explanatory power. *Philos. Sci.* 78, 813–829
60. Pacer, M. *et al.* (2013) Evaluating computational models of explanation using human judgments. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence* (Nicholson, A. and Smyth, P., eds), pp. 498–507, Oregon, AUAI Press Corvallis
61. Wittwer, J. and Renkl, A. (2008) Why instructional explanations often do not work: a framework for understanding the effectiveness of instructional explanations. *Educ. Psychol.* 43, 49–64
62. Sandoval, W.A. and Millwood, K.A. (2005) The quality of students' use of evidence in written scientific explanations. *Cogn. Instr.* 23, 23–55
63. Chi, M.T.H. *et al.* (1989) Self-explanations: how students study and use examples in learning to solve problems. *Cogn. Sci.* 13, 145–182
64. Chi, M.T.H. *et al.* (1994) Eliciting self-explanations improves understanding. *Cogn. Sci.* 18, 439–477
65. Fonseca, B. and Chi, M.T.H. (2011) Instruction based on self-explanation. In *Research on Learning and Instruction* (Mayer, R.E. and Alexander, P.A., eds), pp. 296–321, Routledge
66. Sidney, P.G. *et al.* (2015) How do contrasting cases and self-explanation promote learning? Evidence from fraction division. *Learn. Instr.* 40, 29–38
67. Lombrozo, T. 'Learning by thinking' in science and in everyday life. In *The Scientific Imagination* (Godfrey-Smith, P. and Levy, A., eds), Oxford University Press (in press)
68. Chi, M.T.H. (2000) Self-explaining expository texts: the dual processes of generating inferences and repairing mental models. In *Advances in Instructional Psychology* (Glaser, R., ed.), pp. 161–238, Erlbaum

69. Nokes, T.J. *et al.* (2011) Testing the instructional fit hypothesis: the case of self-explanation prompts. *Instr. Sci.* 39, 645–666
70. Lombrozo, T. (2006) The structure and function of explanations. *Trends in Cognitive Sciences* 10, 464–470
71. Rozenblit, L. and Keil, F. (2002) The misunderstood limits of folk science: an illusion of explanatory depth. *Cogn. Sci.* 26, 521–562
72. Alter, A.L. *et al.* (2010) Missing the trees for the forest: a construal level account of the illusion of explanatory depth. *J. Pers. Soc. Psychol.* 99, 436–451
73. Fernbach, P.M. *et al.* (2013) Explanation fiends and foes: how mechanistic detail determines understanding and preference. *J. Consum. Res.* 39, 1115–1131
74. Edwards, B.J. *et al.* (2013) Effects of explanation and comparison on category learning. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (Knauff, M. *et al.*, eds), pp. 406–411, Austin, TX, Cognitive Science Society
75. Wilkenfeld, D.A. and Lombrozo, T. (2015) Inference to the best explanation (IBE) versus explaining for the best inference (EBI). *Sci. Educ.* 24, 1059–1077
76. Trout, J.D. (2008) Seduction without cause: uncovering explanatory neurophilia. *Trends Cogn. Sci.* 12, 281–282
77. Weisberg, D.S. *et al.* (2008) The seductive allure of neuroscience explanations. *J. Cogn. Neurosci.* 20, 470–477
78. Hopkins, E.J. *et al.* (2016) The seductive allure is a reductive allure: people prefer scientific explanations that contain logically irrelevant reductive information. *Cognition* 155, 67–76
79. Eriksson, K. (2012) The nonsense math effect. *Judgm. Decis. Mak.* 7, 746–749
80. Kelemen, D. (1999) Why are rocks pointy? Children's preference for teleological explanations of the natural world. *Dev. Psychol.* 35, 1440–1452
81. Kelemen, D. and Rosset, E. (2009) The human function compunction: teleological explanation in adults. *Cognition* 111, 138–143
82. Kelemen, D. *et al.* (2013) Professional physical scientists display tenacious teleological tendencies: purpose-based reasoning as a cognitive default. *J. Exp. Psychol. Gen.* 142, 1074–1083
83. Shtulman, A. and Lombrozo, T. (2016) Bundles of contradiction: a coexistence view of conceptual change. In *Core Knowledge and Conceptual Change* (Barner, D. and Baron, A.S., eds), pp. 53–72, Oxford University Press
84. Lombrozo, T. *et al.* (2007) Inferring design: evidence of a preference for teleological explanations in patients with Alzheimer's disease. *Psychol. Sci.* 18, 999–1006
85. Schupbach, J.N. and Sprenger, J. (2011) The logic of explanatory power. *Philos. Sci.* 78, 105–127
86. Craver, C. and Tabery, J. (2016) Mechanisms in science. In *The Stanford Encyclopedia of Philosophy* (Zalta, E.N., ed.),
87. Ahn, W. *et al.* (1995) The role of covariation versus mechanism information in causal attribution. *Cognition* 54, 299–352
88. Vasilyeva, N. and Lombrozo, T. (2015) Explanations and causal judgments are differentially sensitive to covariation and mechanism information. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (Noelle, D.C. *et al.*, eds), pp. 2475–2480, Austin, TX, Cognitive Science Society