Trends in Cognitive Sciences

CellPress
REVIEWS

Opinion

From Probability to Consilience: How Explanatory Values Implement Bayesian Reasoning

Zachary Wojtowicz[1] and Simon DeDeo[1,2,*]

Recent work in cognitive science has uncovered a diversity of explanatory values, or dimensions along which we judge explanations as better or worse. We propose a Bayesian account of these values that clarifies their function and shows how they fit together to guide explanation-making. The resulting taxonomy shows that core values from psychology, statistics, and the philosophy of science emerge from a common mathematical framework and provide insight into why people adopt the explanations they do. This framework not only operationalizes the explanatory virtues associated with, for example, scientific argument-making, but also enables us to reinterpret the explanatory vices that drive phenomena such as conspiracy theories, delusions, and extremist ideologies.

Explaining Explanation

Individuals use a variety of general, cross-domain criteria to evaluate the quality of **explanations** (see Glossary). These **explanatory values** appear in early childhood [1–4] and influence our most sophisticated social knowledge-formation processes [5]. Understanding the foundation of these values is a key goal in the psychology of reasoning [6]. However, despite great empirical progress in demonstrating the existence and importance of explanatory values [7], we lack a unified framework that explains their origin and shows how they fit together to guide our beliefs. The diversity of these values also appears to conflict with Bayesian models of cognition, which some have claimed cannot account for the richness of our explanatory judgments [8,9].

This opinion shows how explanatory values can, in fact, emerge from Bayes' rule, either directly, from an algebraic decomposition of its mathematical structure, or indirectly, from the normative considerations that guide its application. We argue that the resulting set of 'atomic' values, which include **co-explanation**, **descriptiveness**, and **simplicity**, capture many of the existing values proposed by psychologists, philosophers, historians of science, and statisticians. A Bayesian framing provides mathematical definitions of explanatory values, shows how they interact to produce evaluations, and enhances our ability to experimentally probe explanation-making. It leads to the insight that some observed values emerge as different ways for approximating simplicity in practice. It also predicts the existence of new values, and, taking the example of **consilience** [10], shows how complex explanatory values in the history of science can be decomposed into these atoms.

The diverse values seen in the laboratory are often hard to reconcile with each other. People prefer broad explanations that can account for more phenomena [11,12], but seemingly only when those phenomena are actually observed [13]. People generally seem to value simplicity, but the concept has been difficult to pin down; while some studies have demonstrated a preference for **parsimony** [14], others have indicated that this preference may interact with explanatory domain in nontrivial ways [15,16].

Highlights

Recent experiments show that we value explanations for many reasons, such as predictive power and simplicity.

Bayesian rational analysis provides a functional account of these values, along with concrete definitions that allow us to measure and compare them across a variety of contexts, including visual perception, politics, and science.

These values include descriptiveness, co-explanation, and measures of simplicity such as parsimony and concision. The first two are associated with the evaluation of explanations in the light of experience, while the latter concern the intrinsic features of an explanation.

Failures to explain well can be understood as imbalances in these values: a conspiracy theorist, for example, may over-rate co-explanation relative to simplicity, and many similar 'failures to explain' that we see in social life may be analyzable at this level.

[1]Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA, USA
[2]Santa Fe Institute, Santa Fe, NM, USA

*Correspondence:
sdedeo@andrew.cmu.edu (S. DeDeo).

Viewing explanatory values in terms of the computational goal of Bayesian inference resolves some of these conceptual puzzles. This approach, sometimes known as rational analysis [17,18], has been used to explain a wide range of subjective states (e.g., judgments of representativeness [19], suspicious coincidence [20], and subjective randomness [21]) by showing how they help individuals approximate Bayesian thinking. While resource bounds on cognition limit the degree to which the mind actually achieves the Bayesian standard, rational analysis nevertheless provides insights into the origin, function, and form of these mental states. The clarity of the Bayesian framework can also help us understand violations of normative principles, such as judgment biases, conspiracy-mindedness, and delusion.

A Bayesian Framework for Explanatory Values

Probabilistic models of cognition state that people evaluate explanations in terms of how likely they are to be true. Bayesian models of cognition further assume that individuals split this evaluation into two parts: (i) the **log-likelihood**, which measures how probable an explanation makes the observed evidence; and (ii) the **log-prior**, which measures how probable the explanation is independent of the evidence.

We propose that individuals evaluate explanations in terms of the atomic representations that emerge when this decomposition is taken one step further (Box 1). Our framework breaks the log-likelihood into two **empirical values**, descriptiveness and co-explanation, that capture qualitatively different features of how an explanation accounts for evidence.

The log-prior likewise decomposes into two **theoretical values**: a **domain-dependent prior** that captures information such as base-rates, and a domain-general and normatively-grounded simplicity term. As we will show, simplicity is difficult to assess directly. A key feature of our account is how a number of different explanatory values can be understood as heuristics and rules of thumb for approximating it.

Empirical Values: Descriptiveness and Co-explanation

The simplest, or at least 'rough and ready' way to judge an explanation is to consider each piece of evidence independently, tallying the degree to which it makes the explanation look better or worse. This is captured by descriptiveness, the sum of the independent log-probabilities of the relevant facts. This value is familiar from statistics, where it is used to judge a model's fit under the assumption that observations are independent of each other. Although descriptiveness neglects the fact that different pieces of evidence are rarely, if ever, independent, it often works quite well. For example, when evaluating students on the basis of their grades, we can often interpret each grade as an independent reflection of academic ability, thus making grade point average (GPA) a useful summary.

Descriptiveness is generally taken to be an uncontroversial value: all other things being equal, good explanations make each piece of evidence seem more probable. This value has its limits, however, and overemphasis on descriptiveness in a domain where correlations really do matter results in a cognitive bias known as correlation neglect [22].

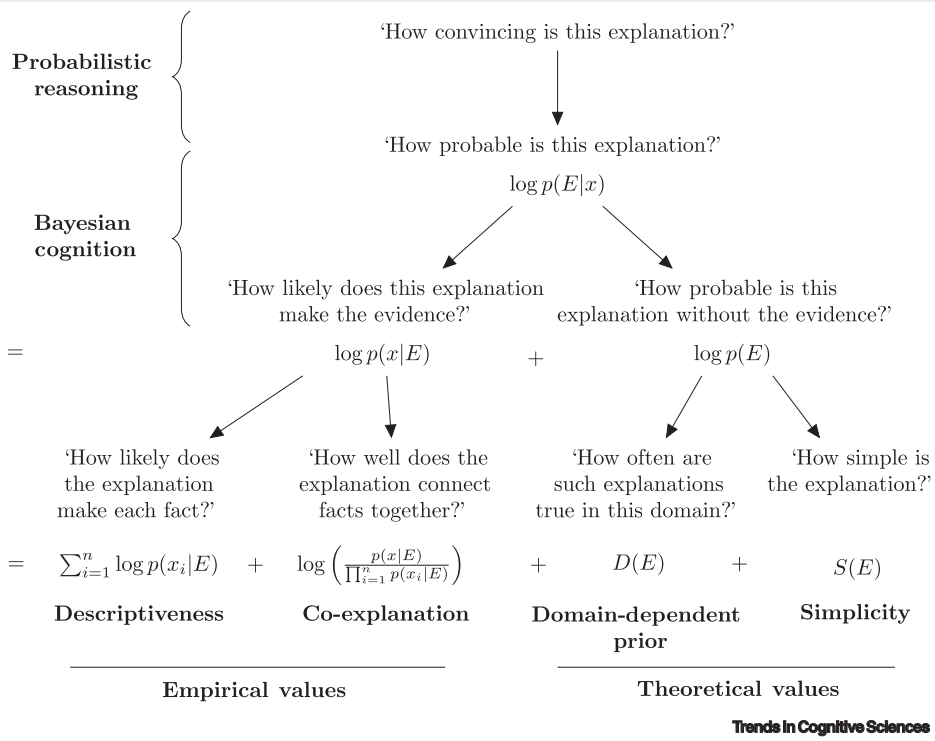
In addition to considering facts in isolation, we also care about how they connect together. This is captured by co-explanation, which measures how well an explanation predicts observations by reference to underlying patterns. Mathematically, co-explanation is the point-wise multi-information between the observations [23] and measures the reduction in uncertainty when one considers observations together instead of separately. The definition follows from the Bayesian decomposition in Box 1 and matches existing proposals for an operationalization of explanatory considerations in the philosophical literature [24,25].

Glossary

Co-explanation: the relative increase in log-probability that an explanation gives a pattern of observed data above its ability to predict each piece in isolation.
Concision: an approximation to simplicity that tracks how briefly or compactly an explanation can be represented within a larger framework.
Consilience: the increased credence one places in an explanation when it fortuitously explains additional phenomena outside its original domain of development.
Descriptiveness: the total log-probability of observed data given an explanation when each observation is considered in isolation.
Domain-dependent prior: a term that captures the influence of domain-specific tacit knowledge and background information (such as base rates).
Empirical values: ways in which an explanation can be valued on the basis of data.
Explanation: an account of some facts about the world. In the Bayesian framework, an explanation supplies a probability distribution over events.
Explanatory values: features of explanations that lead us to prefer one over another.
Log-likelihood: the log-probability of observed data given an explanation.
Log-prior: the log-probability of an explanation before evidence was seen.
Parsimony: an approximation to simplicity that implements some version of Occam's razor. Examples include counting the number of causes or parameters, as well as more formal principles such as the Bayesian Occam's razor.
Simplicity: an umbrella term that captures how 'easily' an explanation may be represented.
Theoretical values: 'priors', or ways in which an explanation can be valued without reference to data.
Unification: the expected co-explanation of data conditional on the explanation being true. Also equal to the mutual information in the case of two variables, or the multi-information in the general case. Measures the degree to which an explanation predicts patterns of outcomes and connects multiple variables together. Another approximation to simplicity.

Box 1. Bayes' Rule Decomposes into Explanatory Values

The cognitive evaluation of an explanation E on the basis of evidence $x = \{x_1, \ldots, x_n\}$ can be broken down into atomic values that collectively decompose the log-additive version of Bayes rule as follows:



Probabilistic models of cognition represent the degrees of belief in an explanation as the (subjective) probability that it is true. Bayesian theories of cognition assert, in turn, that people reason according to Bayes' rule, which implies that these beliefs decompose into two pieces: the likelihood, or the extent to which the explanation makes the things one sees more likely, and the prior, or extent to which one is predisposed to believe the explanation before seeing evidence. Since we are concerned with the relative degrees of belief in different explanations, we omit the normalization constant given that it is constant across the set of explanations for a given body of evidence.

Just as, mathematically, Bayes's rule uses the algebra of probabilities to decompose a degree of belief into a likelihood and a prior, these two objects can, in turn, be further decomposed into terms which, we propose, correspond to values we use to judge explanations. The formal validity of the decomposition of the likelihood into descriptiveness and co-explanation follows from the fact that log$A$ is equal to log$B$ + log A/B. This simple algebraic change has a key cognitive implication, however, as it separates the evaluation of an explanation's relationship to evidence into two intuitively familiar concepts: descriptiveness, which corresponds to the surprise [83] of seeing the items of evidence separately, and co-explanation, which captures the relative surprise of seeing the evidence together, as a whole, versus separately.

In the same way, the log-prior is decomposed into a sum of terms that that each capture considerations that make an explanation more or likely before evidence. Two terms appear: a 'domain' prior that measures the extent the explanation (or ones like it) tends to be true in that domain, and simplicity. Simplicity, as we show, is approximated in a number of different ways that match key features tracked in laboratory experiments.

Validity: increased credence in an explanation caused by its ability to account for 'new' (out-of-sample) evidence.

Co-explanation measures how well an explanation connects patterns of evidence into a whole. It was identified under the name coherence by early empirical work on explanations [26]. Consider learning that two coworkers are secretly dating. This revelation is likely to lead to the feeling of a wide array of disparate observations suddenly 'clicking into place'. The facts about each person considered separately (e.g., the weeks they take vacation, the places

they have visited, the films they have seen) all have innocent explanations. The explanation that the pair is a couple has co-explanatory value because it shows that these facts predict one another: when one of them schedules vacation, we now have reason to suspect the other will as well; if one of them has seen a film, the other one may well have too (since it is likely they saw it together), and so on.

Co-explanation is part of the pleasure of many leisure activities, such as reading detective novels, which often involve the protagonist co-explaining a number of seemingly innocuous facts ('clues', such as 'the dog that did not bark') by identifying the murderer, method, and motive. Equivalently, co-explanation implies mutual predictability: an explanation with high co-explanation suggests that some of the observations could have been be used to predict the others, making the universe seem less arbitrary. The 'eureka' moment when one thinks up a new co-explanatory theory can produce a powerful hedonic response [27] and seems to play an important role in an innate drive for 'sense-making' [28]. These experiences can be seductive and, as with descriptiveness, can be overvalued; among other things, this may impede learning when the patterns to be learned have exceptions [29].

Co-explanation is agnostic on the reasons for predictability. An explanation with co-explanatory power might say that one feature directly causes the other, that features are generated by a hidden common cause, or indeed, that the associations between features emerges 'just so'. While evidence suggests that people do have preferences for certain types of explanations (e.g., causal accounts and, within those, sparse networks that conform to folk intuitions [30–32]) the Bayesian framework requires that such preferences inform theoretical values that are considered separately from the evidence at hand.

That said, invoking common causes is a ubiquitous way to achieve co-explanation in practice and plays a prominent role in domains such as medical diagnosis [33], legal trials [34], and social interactions [26]. Psychological studies of explanation in these domains are often implicit tests of an individual's sensitivity to co-explanation (Box 2). For example, one study [35] trained participants on cases that noted the presence or absence of various symptoms for patients with a fictitious disease. When asked to judge which of two new patients was more likely to have the disease, subjects were sensitive to not only whether each of their symptoms was likely (descriptiveness), but also whether their presentation preserved correlations between symptoms seen in the training set (co-explanation).

Theoretical Values: Domain-Dependent Priors and Simplicity

Theoretical values come in two forms. Any explanation will, in the first instance, be judged in part on the basis of background knowledge. In the Bayesian formulation, such knowledge takes the form of domain-dependent priors. Having well-calibrated priors is part of real-world competence: a good automobile mechanic can anticipate the most likely explanation for an engine failure in a particular model by drawing on their experience with similar cars. The existence of domain-dependent priors is rather uncontroversial and recent investigations have confirmed their influence on explanatory preferences. For example, a subject's diagnostic preferences track their perception of the base-rates of the different diseases [14].

Domain-dependent priors, however, may not be enough [36]. A key insight of what is sometimes called the 'objective' Bayesian perspective is that, even when combined with descriptiveness and co-explanation, domain-specific priors are rarely sufficient for identifying good explanations. This is because one can often improve an explanation (i.e., make it more consistent with what one has observed) by introducing additional complications. Without some counterbalancing force, this will

Box 2. Explanatory Values in Action

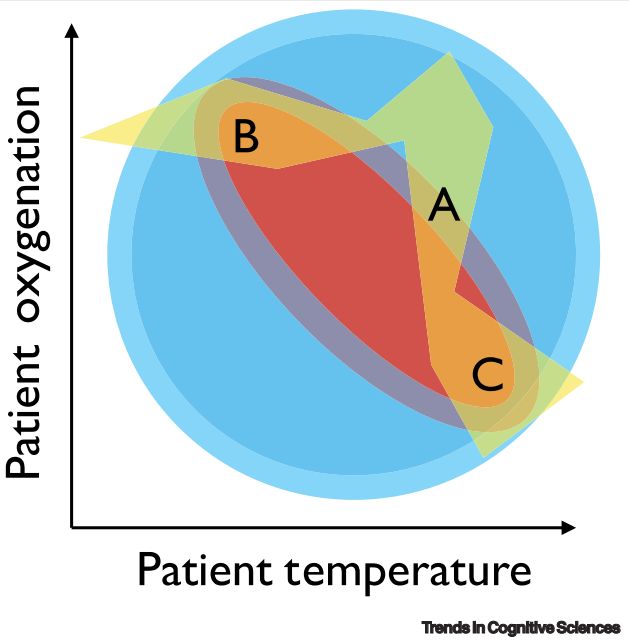


Figure I. Three Diagnoses with Different Explanatory Values.

A common paradigm to tease apart explanatory values is disease diagnosis [12,14,35]. Participants are asked to explain a patient's symptoms by reference to different medical conditions. Figure I illustrates a general case of this task, where there are three potential explanations (shaded in red, blue, and yellow, with density of each color indicating probability) that might produce different patterns of symptoms (here, different combinations of a patient's blood oxygenation and temperature).

In our framework, the blue explanation has low power (it allows for a wide range of outcomes) and low unification (patient temperature is not particularly well predicted by oxygenation). The red explanation has higher power (a narrower range of possibilities) and non-zero co-explanation (high patient temperatures are usually accompanied by low blood oxygenation). The yellow explanation is similar to the red in that it is both powerful and co-explanatory, but has, by contrast, a less simple relationship between oxygenation and temperature.

Confronted with three different patients (A, B, and C), these explanations also have different empirical values. For example, the red explanation has lower descriptiveness than the yellow explanation (patient A falls somewhat outside the normal range for red), but higher co-explanation than the blue (the temperatures of all three patients predict their oxygenation). For the particular case of patient A, yellow also has higher co-explanation than red; under the yellow explanation, a patient with A's oxygenation has a more predictable temperature than under the red one, where the allowable range of temperatures at that oxygenation level is wider.

Which explanation is best depends on context. Even if the yellow explanation is more valued on the basis of descriptiveness and co-explanation, power, or unification, a person may still come to prefer the red, or even the blue, with a strong enough preference for simplicity. For example, the yellow explanation may produce its complex relationship between oxygenation and temperature by invoking the presence of two diseases simultaneously, or through a complicated interaction of different underlying conditions.

Explanations are evaluated relative to a (usually stable) background ontology: here, oxygenation and temperature. If the ontology changes, so do the values and if, for example, doctors worked with a quantity equal to 'temperature minus oxygenation', then the red explanation would become less co-explanatory and more descriptive, while their sum remained constant.

lead to the construction of complicated explanations that account for every detail, even those properly attributed to chance. A variety of normative arguments in probability theory show how a preference for simplicity corrects this tendency [37–39].

In ordinary use there are many answers to what counts as simple. One common conception is parsimony: the ability to explain with fewer 'parts'. Depending on context, these might be constructs, causes, relationships, or parameters. Parsimony is most famously associated with Occam's razor, which states that 'entities should not be multiplied without necessity' [40].

There is good empirical evidence that humans value parsimony. For example, one study [14] found that explanatory preferences were consistent with a Bayesian judgment that included a data-independent prior penalty in favor of parsimony, operationalized as the number of diseases required to explain an alien's symptoms. Further research has shown that such preferences form early in childhood and are robust across contexts [1,4,7,41].

However, individuals sometimes prefer explanations that have more causes [15]. This has been explained as an instance of 'complexity matching', which holds that people prefer explanations that are as complex as the phenomena in question [16] (J. Lim, PhD thesis, University of California, Los Angeles, 2018). The apparent tension between complexity matching and simplicity was resolved by further work [42] confirming (as suggested by [15]) that a preference for complex explanations goes hand in hand with perceptions of their empirical value. A preference for complexity is consistent with a prior for parsimony because subjects use an 'opponent heuristic' that complex (i.e., multiple-cause) explanations typically increase the likelihood of the evidence.

Parsimony is more than a human quirk. It also emerges as a normative principle in Bayesian statistics, where the most straightforward approach is to be parsimonious in the number of 'free parameters': features of an explanation that must be specified before the explanation makes definite predictions. This perspective judges parsimony by reference to an intermediate layer that lies between an explanation and what it is intended to explain. For example, if I explain that my friend is late because their bus broke down, I leave ambiguous the question of precisely where this breakdown occurred. The unspecified detail matters, however: if the bus broke down only a block away, it would not explain why they were late. Informally, 'where the bus broke down' is a free parameter in the explanation.

Two common methods enforce parsimony by penalizing the likelihood: the Akaike Information Criterion (AIC) [43] and the Bayesian Information Criterion (BIC) [44]. For example, in AIC, an explanation must make the observed data roughly 2.7 times more likely in order to compensate for adding an additional parameter.

AIC and BIC are imperfect, however, because they simply count parameters while, intuitively, we judge parsimony based only on the parameters that matter (i.e., those that might affect the explanandum). A generalization of BIC, the Bayesian Occam factor (BOF) [45], shows that the intuition is grounded: some parameters do not affect the complexity of an explanation because they are irrelevant and, whatever value they take, the explanation still provides a good empirical account of the evidence. Other parameters do matter and when the data comes in they must be specified to a greater or lesser degree of rigor. Both sides appear to be in play: people are sensitive to fine-tuning in the free parameters that matter [46], but do not penalize the introduction of irrelevant details [47]. [There is a subtle difference between the BOF and the more general Bayesian Occam's razor (BOR) principle [46]. The BOF approximates BOR, penalizing an explanation on the basis of how fine-tuned the 'best' choice of parameters is; BIC, it can be shown, in turn approximates BOF.]

A feature of BOF is that, in certain situations, the perceived parsimony of an explanation can depend on the particular evidence to hand. This predicts the intriguing possibility that we may

be sensitive to 'revealed' complexity when we judge an explanation. Things can happen that make unexpected demands on the parameters, causing the explanation to look more complicated and fine-tuned in retrospect: imagine discovering that the explanation, earlier, of why a friend was late requires a very precise timing for when the bus broke down. Evidence [48] suggests that some of these *post hoc* considerations tracked by BOF may in fact be captured by the values of descriptiveness and co-explanation. Further work needs to be done to determine how this terminology can be mapped onto the phenomenology of human cognition, an issue we turn to now.

From Parsimony to Concision

While concepts such as BOF provide a rigorous way to define simplicity, they are not the final answer to the problem. One issue is that while parameters have a clear meaning in statistics, they are harder to spot in the informal language of ordinary explanation. Alternative approaches operationalize simplicity in other, potentially more cognitively relevant ways.

For example, Minimum Description Length (MDL) shows that BOF is equivalent to a measure of **concision** (i.e., how succinctly the ambiguities in an explanation can be resolved) [49,50]. According to this perspective, instead of evaluating simplicity in terms of the likelihood of the free parameters (the BOF prescription), one can just as well use the concision of an explanation's expression or perhaps its mental representation [51]. This concision is a rather abstract and 'optimal' one, however, and more work needs to be done to show how it might correspond to, for example, how briefly explanations can be expressed in natural language and how they are encoded by the mind.

Another issue is that any explanation can be understood as a specific instance of a larger, more abstract framework: a disease diagnosis, for example, occurs within a larger causal ontology [32]. This hierarchical nature of explanation [52] means that the concision of an explanation depends upon how far up the hierarchy one looks; an explanation may be very concise within a framework, but the framework itself may be very complicated. It also complicates comparison: to compare the concision of, say, an explanation for poverty in terms of expected utility theory with one in Marxist theory, one would have to estimate, and compare, the concision of the two theories themselves.

These problems are sometimes addressed with 'algorithmic' criteria such as Kolmogorov Complexity [53] and Solomonoff Induction [37]. Theoretically, these measure the 'true' simplicity of any explanation, but technical aspects complicate their use as literal models of cognition. Most notably, both are uncomputable, meaning that it is impossible, in principle, to do better than place upper bounds on concision [54]. An explanation that seems long-winded may have an unexpectedly concise rephrasing and no decision principle can be guaranteed to select the most 'truly' concise explanation from a set.

In the absence of a general method for measuring concision, we fall back on heuristics such as counting causes. One version is **unification**, the expected co-explanation of the explanation conditional upon its truth. Unification says that the world is characterized by patterns, not coincidences. It is a common value in the philosophy of science and some have argued that a good scientific theory makes the manifestation of different phenomena dependent on each other [55] or forms a systemic picture of the order of nature [56].

Causal theories, in particular, are often very good at inducing unification, because they correlate variables with hidden common causes [57]. Such accounts often involve talk in terms of

counterfactuals and not just the probabilities that form the basis of our values here [58]. In many cases, however, such as Pearl's 'do' operator [59], while counterfactual interventions are necessary for causal inference (i.e., for coming to know causes), they are not part of the representation of the resulting explanation itself, which can be done entirely in terms of Bayesian causal nets [60]. We may have to think counterfactually in order choose the right interventions, but this does not mean we require counterfactuals to evaluate the explanations of the outcomes.

Unification tracks concision because it tends to enforce parsimony at the level of unexplained ('root' [61]) causes. A recent study [61] found a preference for root-cause parsimony by showing participants preferred disease diagnoses with fewer unexplained causes, rather than the total number of causes; in the language of Figure 1, an 'elaborate unified' explanation may be more concise than a 'parsimonious disjointed' one. While unification correlates with root-cause parsimony, it is not identical. Unification can be high when a theory has a causal bottleneck [i.e., where a diversity of hidden root causes combine to produce a single hidden cause that then explains (and correlates) the evidence]. A theory that explains symptoms by reference to a single disease has both high unification and parsimony in root causes. Another theory that includes a complex, multicausal account of how the disease itself came to be contracted has equal unification, but more root causes.

Another source of concision is 'uniformity' [16], or the consistency of relationships among the details of an explanation. Uniformity may provide concision even when parsimony is lacking because many causes may be described at once. It is concise, for example, to say that a traffic jam was caused by everyone trying to get to work at the same time. There may be hundreds of separate causes (the desires of each driver to get to work), but they are all, in effect, the same. Concision may also lead to a preference for explanations that show how new phenomena can be put into analogy with familiar ones [62]. A successful analogy means that one can express the large

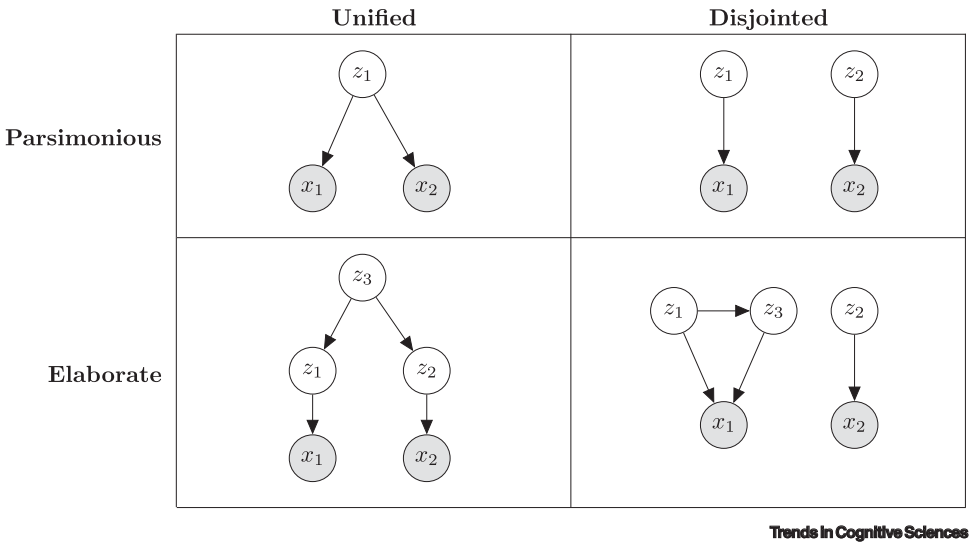


Figure 1. Diagrams Representing the Causal Structure of Different Explanations. Shaded nodes represent observables, other nodes represent postulated latent causes, and arrows represent causal relationships. One way of assessing simplicity is parsimony, which counts the number of parameters, or latent causes invoked by an explanation. A second way of assessing simplicity is unification, which measures the degree to which a theory provides an overarching, connected account of multiple features of the world. Explanations can vary along each of these dimensions independently, so that their overall 'simplicity' might be judged as a weighted combination of the two.

number of relationships necessary to achieve descriptiveness and co-explanation very concisely by saying they are just like something already known, along with a relatively brief statement of what corresponds to what.

Validity and the Virtue of Not Overfitting

If simplicity is not taken into account, one generally produces 'overfit' explanations with complications that enable inappropriate fine-tuning to match the data. An infamous example of overfitting in statistics is *p*-hacking. Overfit explanations work 'too well': they explain not only the law-like regularities in the data, but also coincidences and noise. This makes them fragile, because they latch on to misleading coincidences to predict what will happen next.

The very fragility of an overfit explanation, however, suggests a way to avoid it: before you build your explanations, hold some data in reserve. After building them, compare their empirical values solely on this held-out data; overfit explanations will underperform. In statistics, this method is known as cross-validation [63].

The power of cross-validation suggests, in turn, the possibility of a corresponding value in human cognition. We call this **validity**: the value ascribed to an explanation when it is able to predict new evidence. In the ordinary course of life, it is difficult to deliberately conceal knowledge from one's self while constructing explanations and it is (at the very least) impolite to do so to someone else. It does happen naturally, however, when we uncover new information either actively or passively. If validity is a value, this newly manifested information should have greater impact on our credence in an explanation than the data that came before.

While validity has not been the subject of detailed empirical study in the psychology of explanation-making, it appears quite early in the philosophical literature as part of consilience. Consilience was introduced by William Whewell [64] in the 19th century to describe features of scientific explanations that, he argued, both are, and ought to be, prized by the community. As defined by Whewell, 'the Consilience of Inductions takes place when an induction, obtained from one class of facts, coincides with an induction, obtained from another class' (i.e., when an explanation constructed to explain phenomena in one domain turns out to predict phenomena in an entirely different one as well). Importantly, these new facts were not considered during its construction. For Whewell, 'such a coincidence of untried facts with speculative assertions cannot be the work of chance, but implies some portion of truth in the principles on which the reasoning is founded'.

Consilience is an example of how values combine: in this case, the values of concision, unification, and validity. A consilient explanation is certainly more concise than what we had before: at the very least, it can replace whatever previous explanation we had for the facts of the second domain without becoming more complex. It is also more unified, making the second domain dependent on the same things as the first. Finally, that the facts of the second domain are 'untried' gives a consilient explanation greater validity.

Other Virtues, Other Vices

A prominent alternative to Bayesian conceptions of explanation is inference to the best explanation (IBE) [5], which says belief formation is, or should be, guided by 'explanatory considerations' [7,65]. In Harman's original formulation, these were additional factors that justified accepting one hypothesis from a pool of alternatives. While IBE has traditionally been opposed to Bayesianism, it has also been somewhat loosely defined, making it possible for some to argue that the two are either compatible [66–70] or potentially even identical [71,72].

Recent work on 'explanationism' [8,9] has sharpened and revived the tension. Explanationism is the hypothesis that people update explanations in a way that violates Bayes' rule and uses, instead, a separate group of IBE-like, non-Bayesian explanatory values. Depending on one's goals, these values may even be improvements on Bayes' rule, allowing one to get the right answer, more quickly, most of the time [8]. Intriguingly, a pair of recent papers [73,74] that reanalyzes earlier data [75] finds evidence for two such rules, associated with Karl Popper and I.J. Good.

Because explanationism replaces Bayes' rule, it revises the decomposition in Box 1. In proposal [8], this is equivalent to adding a new term to the decomposition: a boost (or penalty) that combines both data and intrinsic properties of the explanation, leading to a correction factor that can be tested in the laboratory. Explanationism is incompatible with Bayesianism, but there is a clear path forward for separating them experimentally.

Indeed, one of the benefits of our decomposition is that it allows us to quantify deviations from Bayesian behavior due not only to potential non-Bayesian virtues such as IBE, but also to 'vices' of thought caused by improper weighting of the terms in play (Box 3). Deviations, for example, that overweight co-explanation, correspond to undue preference for correlations. Other sources of non-Bayesian behavior may come from cognitive constraints: this appears to be the explanation for 'latent scope bias' [13], which appears to derive from the use of an

Box 3. When Values Become Vices

One benefit of our framework is that it enables us to understand characteristic explanatory pathologies, the subject of 'vice epistemology' [84], as deviations from the normative weighting of explanatory values.

Consider the phenomenon of overgeneralization [29] (i.e., attempting to cover all examples with a single explanation rather than allowing for exceptions). This can be caused by over-valuing co-explanation relative to descriptiveness, or by over-valuing unification (since theories that have high unification will tend to have higher co-explanation when they are good fits to the data). Recent empirical work has started to tie abnormal reasoning to common inferential biases that generalize across domains in a way that suggests the systemic miscalibration of values may be at fault.

For example, those prone to paranormal thinking also show greater susceptibility to the conjunction fallacy [85], which can result from overvaluing co-explanation because labeling Linda a feminist as well as a bank teller [86] provides a co-explanatory account of her behavior. There are also strong individual differences in the tendency to believe conspiracies: those who believe one are more likely to believe others [87]. Notably, this trait is common in individuals with schizotypal disorder [88], which is in turn linked to many other explanatory abnormalities [89].

Conspiracy theories are often both abnormally co-explanatory and descriptive [90]. They account for anomalous facts that are unlikely under the 'official' explanation ('errant data' [91], as exemplified by, e.g., Oklahoma City bombing conspiracy theories [92]) and show how seemingly arbitrary facts of ordinary life are correlated by hidden events [93]. Along these lines, manipulations that induce subjects to see illusory correlations in neutral domains, like stock returns, also increase beliefs in conspiracy theories [94]. Finally, and famously, conspiracy theories are unifying: they describe a universe where everything is correlated by a network of hidden common causes: the motives and meetings of the conspirators [95].

Valuing these features is not, in and of itself, a vice; what frequently goes wrong is the failure to balance them against others. On the surface, a conspiracy theory is simple; as it is unfolded, however, increasing complexity is required to explain contradictory evidence and the cover-up that has, so far, prevented it from coming to light. Such a judgment is itself open to criticism; as noted by [96], some conspiracies are extremely compelling on normative grounds. Some even turn out to be true.

Striking a virtuous balance between so many considerations is itself a challenging cognitive problem, one that we solve partially by reliance on the judgments of others to correct our faults. Reliance on the faulty values of others might help explain membership in antivaccination movements [97], COVID-19 conspiracies [98], the use of pseudoscience in extremist ideologies [99], and science denialism [100]. However, while these beliefs are in part formed and maintained by social processes, they interact with individual-level predispositions. One avenue for future research is how social processes serve to maintain, accentuate, or exploit individual-level explanatory miscalibrations.

improper heuristic ('inferred evidence') for the otherwise Bayesian evaluation of empirical values in terms of 'manifest scope' [76,77].

Concluding Remarks and Future Directions

This opinion article suggests that it may be possible to explain the diversity of how we judge and compare explanations using a small number of atomic values. It says that parsimony, concision, and unification, for example, are different forms of an underlying simplicity value, and that we are sensitive to two forms of empirical value, descriptiveness and co-explanation. It predicts the existence of new values, such as revealed complexity and validity, that correspond to features of Bayesian statistical inference. In other cases, it can help show how explanatory values in the history of science, such as consilience, are complex combinations of these atoms. Our claims are supported in part by research that shows that preferences that appear to be incompatible with these values, such as complexity matching and narrow latent scope, may in fact be consequences of heuristics, rather than separate values in their own right.

We have focused on how explanatory values influence our judgments of explanations in the presence of data. This is not all values are called on to do, however, and a key question is how values influence the way people construct explanations [78] and test and reason about them in a social context [79]. A recent study [80], for example, provides an account of scientific investigation where decidedly non-Bayesian criteria, such as attention to salient outliers, guide which explanations to explore and modify next. Explanatory values are only partial guides to judgments of what stands in need of an explanation [81] and non-epistemic values, such as future utility, influence how satisfied we are when we get them [82]. We expand on a number of other exciting avenues for future work (see Outstanding Questions).

Acknowledgments

We acknowledge the support of the John Templeton Foundation and Jaan Tallinn via the Survival and Flourishing Fund. We thank Colin Allen, Helena Miton, Danny Oppenheimer, Robert X.D. Hawkins, and Reza Negarestani for their insightful comments on earlier drafts of this work.

References

1. Walker, C.M. *et al.* (2017) Effects of explaining on children's preference for simpler hypotheses. *Psych. Bull. Rev.* 24, 1538–1547
2. Samarapungavan, A. (1992) Children's judgements in theory choice tasks: scientific rationality in childhood. *Cognition* 45, 1–32
3. Frazier, B.N. *et al.* (2016) Young children prefer and remember satisfying explanations. *J. Cogn. Dev.* 17, 718–736
4. Bonawitz, E.B. and Lombrozo, T. (2012) Occam's rattle: children's use of simplicity and probability to constrain inference. *Dev. Psychol.* 48, 1156
5. Harman, G.H. (1965) The inference to the best explanation. *Philos. Rev.* 74, 88–95
6. Oaksford, M. and Chater, N. (2020) New paradigms in the psychology of reasoning. *Annu. Rev. Psychol.* 71
7. Lombrozo, T. (2016) Explanatory preferences shape learning and inference. *Trends Cogn. Sci.* 20, 748–759
8. Douven, I. and Wenmackers, S. (2017) Inference to the best explanation versus Bayes' rule in a social setting. *Br. J. Philos. Sci.* 68, 535–570
9. Douven, I. (2017) Inference to the best explanation: what is it? And why should we care. In *Best Explanations: New Essays on Inference to the Best Explanation* (McCain, K. and Poston, T., eds), pp. 4–22, Oxford University Press
10. Laudan, L. (1971) William Whewell on the consilience of inductions. *Monist* 55, 368–391
11. Rebitschek, F.G. *et al.* (2016) The diversity effect in diagnostic reasoning. *Mem. Cogn.* 44, 789–805
12. Johnson, S. *et al.* (2014) Explanatory scope informs causal strength inferences. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol 36)
13. Khemlani, S.S. *et al.* (2011) Harry Potter and the sorcerer's scope: latent scope biases in explanatory reasoning. *Mem. Cogn.* 39, 527–535
14. Lombrozo, T. (2007) Simplicity and probability in causal explanation. *Cogn. Psychol.* 55, 232–257
15. Zemla, J.C. *et al.* (2017) Evaluating everyday explanations. *Psychon. Bull. Rev.* 24, 1488–1500
16. Lim, J.B. and Oppenheimer, D.M. (2020) Explanatory preferences for complexity matching. *PLoS One* 15, e0230929
17. Lieder, F. and Griffiths, T.L. (2019) Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behav. Brain Sci.* 43, e1
18. Chater, N. and Oaksford, M. (1999) Ten years of the rational analysis of cognition. *Trends Cogn. Sci.* 3, 57–65
19. Tenenbaum, J.B. *et al.* (2001) The rational basis of representativeness. In *Proceedings of the 23rd annual conference of the Cognitive Science Society*, pp. 1036–1041, Citeseer
20. Griffiths, T.L. and Tenenbaum, J.B. (2007) From mere coincidences to meaningful discoveries. *Cognition* 103, 180–226
21. Griffiths, T.L. *et al.* (2018) Subjective randomness as statistical inference. *Cogn. Psychol.* 103, 85–109
22. Enke, B. and Zimmermann, F. (2019) Correlation neglect in belief formation. *Rev. Econ. Stud.* 86, 313–332
23. Studený, M. and Vejnarová, J. (1998) The multiinformation function as a tool for measuring stochastic dependence. In *Learning in graphical models* (Jordan, M., ed.), pp. 261–297, Springer
24. Myrvold, W.C. (2003) A Bayesian account of the virtue of unification. *Philos. Sci.* 70, 399–423

Outstanding Questions

How are explanatory values calculated and represented by the mind?

What is the phenomenology of explanatory values? Do different values have a distinct 'feel' when we evaluate explanations? What is the relationship between these values and other epistemic states, such as curiosity?

To what extent are values determined during early childhood development, versus learned in later life? Can people change their values in response to experience or teaching?

What social, cultural, or psychological forces lead to the creation of more complex 'molecular' values such as consilience?

How do explanatory values influence the cultural evolution of explanations?

What determines the categories (i.e., variables) over which explanatory values are evaluated? How do these coevolve with explanations? Are these categories determined by other forces, or are they, at least partially, determined by explanatory considerations themselves?

How universal are these values? How much of the difference between individual preferences for explanations is driven by domain-general explanatory values versus contextual priors?

What is the connection between organic brain diseases and imbalanced explanatory values? What can this tell us about the neurological basis of these values and the manner in which they are assessed?

To what extent are social movements associated with pathological beliefs, such as conspiracy theories, driven by explanatory imbalance? Does participation in such a movement reinforce such imbalances?

To what extent are values simply a means of achieving the practical goal of prediction? What other roles do they play in human life?

What counts as 'an explanation' in the first place? What features make a

25. Zhang, J. and Zhang, K. (2015) Likelihood and consilience: on Forster's counterexamples to the likelihood theory of evidence. *Philos. Sci.* 82, 930–940
26. Read, S.J. and Marcus-Newhall, A. (1993) Explanatory coherence in social explanations: A parallel distributed processing account. *J. Pers. Soc. Psychol.* 65, 429
27. Gopnik, A. (1998) Explanation as orgasm. *Mind. Mach.* 8, 101–118
28. Chater, N. and Loewenstein, G. (2016) The under-appreciated drive for sense-making. *J. Econ. Behav. Organ.* 126, 137–154
29. Williams, J.J. *et al.* (2013) The hazards of explanation: overgeneralization in the face of exceptions. *J. Exp. Psychol. Gen.* 142, 1006
30. Lu, H. *et al.* (2008) Bayesian generic priors for causal learning. *Psychol. Rev.* 115, 955
31. Yeung, S. and Griffiths, T.L. (2015) Identifying expectations about the strength of causal relationships. *Cogn. Psychol.* 76, 1–29
32. Tenenbaum, J.B. *et al.* (2007) Intuitive theories as grammars for causal inference. In *Causal Learning: Psychology, Philosophy, and Computation* (Gopnik, A. and Schulz, L., eds), pp. 201–322, Oxford University Press
33. Dragulinescu, S. (2016) Inference to the best explanation and mechanisms in medicine. *Theor. Med. Bioeth.* 37, 211–232
34. Amaya, A. (2016) Inference to the best legal explanation. In *Legal Evidence and Proof* (Kaptein, H. *et al.*, eds), pp. 149–174, Routledge
35. Medin, D.L. *et al.* (1982) Correlated symptoms and simulated medical classification. *J. Exp. Psychol. Learn. Mem. Cogn.* 8, 37
36. Goodman, N.D. *et al.* (2011) Learning a theory of causality. *Psychol. Rev.* 118, 110
37. Solomonoff, R.J. (1964) A formal theory of inductive inference. Part I. *Inf. Control.* 7, 1–22
38. Rissanen, J. (1983) A universal prior for integers and estimation by minimum description length. *Ann. Stat.* 416–431
39. Chater, N. and Vitányi, P. (2003) Simplicity: A unifying principle in cognitive science? *Trends Cogn. Sci.* 7, 19–22
40. Baker, A. (2004) Simplicity. In *The Stanford Encyclopedia of Philosophy* (Zalta, E.N., ed.), Stanford University
41. Bonawitz, E.B. and Lombrozo, T. (2007) Simplicity and probability in children's causal explanations. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol 29)
42. Johnson, S.G.B. *et al.* (2019) Simplicity and complexity preferences in causal explanation: An opponent heuristic account. *Cogn. Psychol.* 113, 101222
43. Akaike, H. (1974) A new look at the statistical model identification. *IEEE Trans. Autom. Control* 19, 716–723
44. Schwarz, G. *et al.* (1978) Estimating the dimension of a model. *Ann. Stat.* 6, 461–464
45. MacKay, D.J.C. (2003) *Information Theory, Inference and Learning Algorithms*, Cambridge University Press
46. Blanchard, T. *et al.* (2018) Bayesian Occam's razor is a razor of the people. *Cogn. Sci.* 42, 1345–1359
47. Bechlivanidis, C. *et al.* (2017) Concreteness and abstraction in everyday explanation. *Psychon. Bull. Rev.* 24, 1451–1464
48. Tenenbaum, J.B. and Griffiths, T.L. (2001) Generalization, similarity, and Bayesian inference. *Behav. Brain Sci.* 24, 629
49. Rissanen, J. (1978) Modeling by shortest data description. *Automatica* 14, 465–471
50. MacKay, D.J.C. (1992) Bayesian interpolation. *Neural Comput.* 4, 415–447
51. Chater, N. (1996) Reconciling simplicity and likelihood principles in perceptual organization. *Psychol. Rev.* 103, 566
52. Henderson, L. *et al.* (2010) The structure and dynamics of scientific theories: a hierarchical Bayesian perspective. *Philos. Sci.* 77, 172–200
53. Kolmogorov, A.N. (1965) Three approaches to the quantitative definition of 'information'. *Probl. Inf. Transm.* 1, 1–7
54. Negarestani, R. (2018) *Intelligence and Spirit*, Urbanomic Press
55. Friedman, M. (1974) Explanation and scientific understanding. *J. Philos.* 71, 5–19
56. Kitcher, P. (1989) Explanatory unification and the causal structure of the world. In *Minnesota Studies in the Philosophy of Science* (Love, A.C., ed.), University of Minnesota Press
57. Salmon, W.C. (1998) *Causality and Explanation*, Oxford University Press
58. Halpern, J.Y. and Pearl, J. (2005) Causes and explanations: a structural-model approach. Part II: explanations. *Br. J. Philos. Sci.* 56, 889–911
59. Pearl, J. (2009) *Causality*, Cambridge University Press
60. Tenenbaum, J.B. and Griffiths, T.L. (2003) Theory-based causal inference. In *Advances in Neural Information Processing Systems*, pp. 43–50
61. Pacer, M. and Lombrozo, T. (2017) Ockham's razor cuts to the root: Simplicity in causal explanation. *J. Exp. Psychol. Gen.* 146, 1761
62. Thagard, P.R. (1978) The best explanation: criteria for theory choice. *J. Philos.* 75, 76–92
63. Gelman, A. *et al.* (2014) Understanding predictive information criteria for Bayesian models. *Stat. Comput.* 24, 997–1016
64. Whewell, W. (1847) In *The Philosophy of the Inductive Sciences: Founded Upon Their History* (Vol 2, 2nd edn), pp. 65, John W. Parker
65. Schupbach, J.N. (2016) 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), pp. 39–61, Oxford University Press
66. Okasha, S. (2000) Van Fraassen's critique of inference to the best explanation. *Stud. Hist. Phil. Sci. A* 31, 691–710
67. Lipton, P. (2004) *Inference to the Best Explanation*, Taylor & Francis
68. Huemer, M. (2009) Explanationist aid for the theory of inductive logic. *Br. J. Philos. Sci.* 60, 345–375
69. Weisberg, J. (2009) Locating IBE in the Bayesian framework. *Synthese* 167, 125–143
70. Climenhaga, N. (2017) Inference to the best explanation made incoherent. *J. Philos.* 114, 251–273
71. Van Fraassen, B.C. (1989) *Laws and Symmetry*, Clarendon
72. Henderson, L. (2014) Bayesianism and inference to the best explanation. *Br. J. Philos. Sci.* 65, 687–715
73. Douven, I. and Schupbach, J.N. (2015) Probabilistic alternatives to Bayesianism: the case of explanationism. *Front. Psychol.* 6, 459
74. Douven, I. and Schupbach, J.N. (2015) The role of explanatory considerations in updating. *Cognition* 142, 299–311
75. Schupbach, J.N. (2011) Comparing probabilistic measures of explanatory power. *Philos. Sci.* 78, 813–829
76. Johnson, S.G.B. *et al.* (2016) Sense-making under ignorance. *Cogn. Psychol.* 89, 39–70
77. Johnston, A.M. *et al.* (2017) Little Bayesians or little Einsteins? Probability and explanatory virtue in children's inferences. *Dev. Sci.* 20, e12483
78. Horne, Z. *et al.* (2019) Explanation as a cognitive process. *Trends Cogn. Sci.* 23, 187–199
79. Mercier, H. and Sperber, D. (2017) *The Enigma of Reason*, Harvard University Press
80. Gelman, A. and Shalizi, C.R. (2013) Philosophy and the practice of Bayesian statistics. *Br. J. Math. Stat. Psychol.* 66, 8–38
81. Liquin, E. and Lombrozo, T. (2018) Determinants and consequences of the need for explanation. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, Cognitive Science Society
82. Liquin, E. and Lombrozo, T. (2019) Inquiry, theory-formation, and the phenomenology of explanation. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, pp. 664–670, Cognitive Science Society
83. Itti, L. and Baldi, P. (2009) Bayesian surprise attracts human attention. *Vis. Res.* 49, 1295–1306
84. Cassam, Q. (2016) Vice epistemology. *Monist* 99, 159–180
85. Brotherton, R. and French, C.C. (2014) Belief in conspiracy theories and susceptibility to the conjunction fallacy. *Appl. Cogn. Psychol.* 28, 238–248
86. Tversky, A. and Kahneman, D. (1974) Judgment under uncertainty: heuristics and biases. *Science* 185, 1124–1131
87. Bruder, M. *et al.* (2013) Measuring individual differences in generic beliefs in conspiracy theories across cultures: conspiracy mentality questionnaire. *Front. Psychol.* 4, 225

predictive model into something that is intelligible to us and therefore answerable to our explanatory values?

88. Darwin, H. *et al.* (2011) Belief in conspiracy theories. The role of paranormal belief, paranoid ideation and schizotypy. *Personal. Individ. Differ.* 50, 1289–1293
89. McLean, B.F. *et al.* (2017) Association of the jumping to conclusions and evidence integration biases with delusions in psychosis: a detailed meta-analysis. *Schizophr. Bull.* 43, 344–354
90. Vitriol, J.A. and Marsh, J.K. (2018) The illusion of explanatory depth and endorsement of conspiracy beliefs. *Eur. J. Soc. Psychol.* 48, 955–969
91. Keeley, B.L. (1999) Of conspiracy theories. *J. Philos.* 96, 109–126
92. Coady, D. (1993) In *Conspiracy Theories: The Philosophical Debate* (Vol 4), pp. 201–213, Ashgate Publishing
93. Tangherlini, T. (2017) Toward a generative model of legend: pizzas, bridges, vaccines, and witches. *Humanities* 7, 1
94. Whitson, J.A. and Galinsky, A.D. (2008) Lacking control increases illusory pattern perception. *Science* 322, 115–117
95. Douglas, K.M. *et al.* (2017) The psychology of conspiracy theories. *Curr. Dir. Psychol. Sci.* 26, 538–542
96. Dentith, M.R.X. (2016) When inferring to a conspiracy might be the best explanation. *Soc. Epistemol.* 30, 572–591
97. Miton, H. and Mercier, H. (2015) Cognitive obstacles to pro-vaccination beliefs. *Trends Cogn. Sci.* 19, 633–636
98. Shahsavari, S. *et al.* (2020) Conspiracy in the time of corona: automatic detection of covid-19 conspiracy theories in social media and the news. *Comput. Math.* Published online August 4, 2020. http://doi.org/10.21203/rs.3.rs-52079/v1
99. O'Neill, R. (2018) *Seduction: Men, Masculinity and Mediated Intimacy*, John Wiley & Sons
100. Rutjens, B.T. *et al.* (2018) Attitudes towards science. In *Advances in Experimental Social Psychology* (Vol 57), pp. 125–165, Elsevier