From Probability to Consilience: How Explanatory Values Implement Bayesian Reasoning

Zachary Wojtowicz* Simon DeDeo*†

Abstract

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 how these values fit together to guide explanation. The resulting taxonomy provides a set of predictors for which explanations people prefer and shows how core values from psychology, statistics, and the philosophy of science emerge from a common mathematical framework. In addition to operationalizing the explanatory *virtues* associated with, for example, scientific argument-making, this framework also enables us to reinterpret the explanatory *vices* that drive conspiracy theories, delusions, and extremist ideologies.

I Explaining explanation

Intuitively, philosophically, and as seen in laboratory experiments, **explanations** are judged as better or worse on the basis of many different criteria. These **explanatory values** appear in early childhood [1, 2, 3, 4, 5] and their influence extends to some of the most sophisticated social knowledge formation processes we know [6]. We lack, however, an understanding of the origin of these values or an account of how they fit together to guide belief formation. The multiplicity of values also appears to conflict with Bayesian models of cognition, which speak solely in terms of degrees of beliefs and suggest we judge explanations as better or worse on the basis of a single quantity, the posterior **likelihood** (see Glossary). In this opinion, we show how to resolve these conflicts by arguing that previously-identified explanatory values capture different components of a full Bayesian calculation and, when considered together and weighed appropriately, implement Bayesian cognition.

This framework shows how key explanatory values identified by laboratory experiments and philosophers of science—**co-explanation**, **descriptiveness**, **precision**, **unification**, **power**, and **simplicity**—emerge naturally from the mathematical structure of probabilistic inference, thereby reconciling them with Bayesian models of cognition [7, 8]. Second, it shows how these values combine to produce preferences for one explanation over another. Third, it emphasizes new conceptual distinctions, such as one between explanatory values that can be assessed before the arrival of data (**theoretical values**) and those that can only be assessed after the arrival of data (**empirical values**). Finally, it enables us to reinterpret work on the characteristic deviations from normative patterns of explanation that drive phenomena such as conspiracy theories, delusions, and extremist ideologies.

It also resolves a tension in the influential philosophical account of "inference to the best explanation" (IBE; [1, 6, 9]) which says belief formation is, or should be, guided by explanatory considerations. While some hold that IBE is incompatible with Bayesian updating because explanatory considerations cannot be captured within a probabilistic framework [10, 11], others argue that the two are either compatible [12, 13, 14, 15] or potentially even identical [16, 17]. We adopt this latter perspective, and show how our framework provides a compelling—albeit preliminary—account of how such an "emergent compatabilism" [16] can be achieved. [1]

*Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA
†Santa Fe Institute, Santa Fe, NM

[1] A more nuanced difference between Bayesianism and many variants of IBE is that the former, on its own, provides a system for determining the relative strength of various theories while the latter actually provides grounds for accepting a single, "best" explanation [18]. Clearly, however, one can combine Bayesian updating with any number of rules for acceptance, such as choosing the explanation that has the maximum a posteriori probability (MAP) of being correct.

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, unification, power, and simplicity, and fall into two groups: 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.

Box 1: Bayes' Rule Decomposes into Explanatory Values

Formally, an explanation E is evaluated on the basis of its log-likelihood in the presence of evidence $x = \{x_1, \ldots, x_n\}$; using Bayes rule, this gives us

$$\log p(E|x) = \underbrace{\sum_{i=1}^{n} \log p(x_i|E)}_{\text{Descriptiveness}} + \underbrace{\log\left(\frac{p(x|E)}{\prod_{i=1}^{n} p(x_i|E)}\right)}_{\text{Co-explanation}} + \underbrace{\sum_{i} \log T_i(E)}_{\text{Theoretical Values}} + \underbrace{\log \pi(E)}_{\text{Contextual Factors}}$$

The terms in this decomposition are: (i) descriptiveness, which measures the total extent to which the explanation predicts each fact in isolation from the others; (ii) co-explanation, which measures the extent to which the explanation links facts together; (iii) theoretical, or evidence-independent values; and (iv) context-dependent priors. (For simplicity, we do not show the additional normalization term that is constant across all explanations and thus does not affect comparisons between explanations.)

II A Bayesian Framework for Explanatory Values

Bayes' rule says that we should value an explanation in terms of a "posterior" degree of belief, determined by our prior degree of belief in that explanation times the probability that it assigns to the data we have observed. Working with log-probabilities means that the components of this calculation combine additively, matching the intuition that we weigh multiple features of an explanation by adding them together to make final determinations about its validity.

Box 1 shows how the log-posterior can be rearranged into a series of additive terms which define, mathematically, values identified across a variety of different contexts. The first two terms are empirical, and track how an explanation accounts for observed data: descriptiveness (formally, the log-likelihood under the assumption that the data are independent), measures how well an explanation predicts the facts in isolation, and co-explanation (formally, the information-theoretic pointwise multi-information), measures how well an explanation predicts patterns that connect facts together. The next terms are theoretical and track the value of an explanation independently from the data. As we will argue, two key theoretical values correspond to expected empirical values, while others reflect structural features of an explanation and context-depenent priors.

This leads to two pairs of explanatory values—descriptiveness and power, co-explanation and unification—that appear either at the empirical or the theoretical stage, respectively, along with an additional theoretical value of simplicity. The correspondences between the mathematical terms and the explanatory values are shown in the Glossary. We discuss each in turn.

III Explanation through the Lens of Descriptiveness and Power

The simplest way to judge an explanation is to consider each piece of evidence for it independently, keeping a running tally of the degree to which it makes the explanation look better or worse. This is captured by descriptiveness, the sum of the independent log-probabilites of the relevant facts.[2]

Although descriptiveness neglects that facts are rarely independent, it nevertheless often works quite well. For example, when evaluating students on the basis of their grades, we can usually interpret each mark as an independent reflection of academic ability, thus making GPA a useful summary. On the other hand, overemphasis on descriptiveness in a domain where correlations really do matter results in a cognitive bias known as correlation neglect [20].

The theoretical value corresponding to descriptiveness is power: how descriptive an explanation is in a world where it was true. Valuing power means valuing explanations that make more definite predictions. All other things being equal, more descriptiveness is always a good thing. Power is more ambiguous, and someone might consider power to be a virtue or a vice. High-power explanations make definitive predictions and therefore more easily falsified; they are also more easily learned from experience. Further, if you believe a high power explanation, you expect those who value descriptiveness to be receptive to it as well.

Power can also be a vice, however, because the world is not always as predictable as we might wish. In uncertain situations, one might value low power explanations as more open-minded and allowing for a wider range of possibilities. Indeed, in statistics, Ref. [21] has advocated for the "principle of maximum entropy", which views minimizing **precision**—the sum of power and unification (discussed below)—as a universal normative rule of inference because it presumes to know the least *a priori*.

[2] A similar approximation is frequently employed in statistics—the familiar identically and independently distributed (IID) assumption.

Glossary

Explanation: an account of some observable aspect of the world. In the Bayesian framework, an explanation supplies a probability distribution over events.

Explanatory Values: explanatory features that lead us to prefer one explanation over another.

Empirical Values:

(Ways in which an explanation can be valued on the basis of data.)

Log-Likelihood: $\log p(x|E)$
The log-probability of observed data given an explanation.

Descriptiveness: $\sum_i \log p(x_i|E)$
The total log-probability of observed data given an explanation when each observation is considered in isolation.

Co-explanation: $\log\left(\frac{p(x|E)}{\prod_{i=1}^{n} p(x_i|E)}\right)$
The relative increase in log-probability that an explanation gives a pattern of observed data above its ability to predict each piece in isolation. Equal to point-wise mutual information in the case of two variables, or point-wise multi-information in the general case [19].

Theoretical Values:

("Priors", or ways in which an explanation can be valued without reference to data.)

Precision: $\mathbb{E}_E[\log p(x|E)]$

The expected likelihood of data conditional on the explanation being true. Also equal to the negative entropy of the explanation. Measures the degree to which an explanation's predictions concentrate in a particular subset of the space of all possible outcomes.

Power: $\mathbb{E}_E[\sum_i \log p(x_i|E)]$

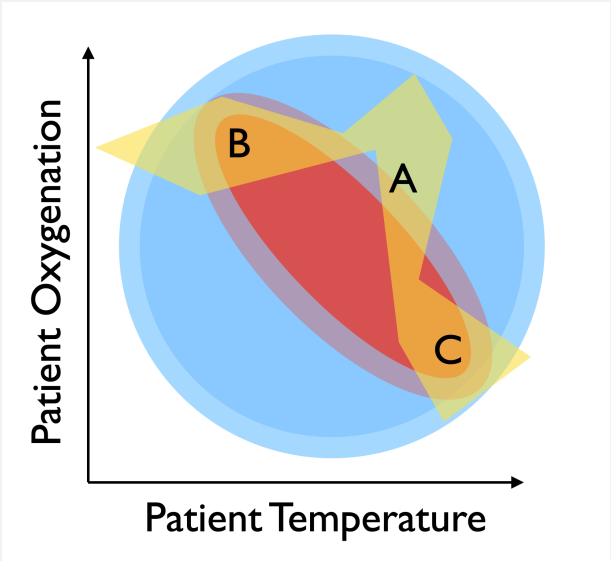
The expected descriptiveness of data conditional on the explanation being true. Measures the degree to which an explanation tends to produce individual pieces of data that it can account for in isolation.

Unification: $\mathbb{E}_E\left[\log\left(\frac{p(x|E)}{\prod_{i=1}^{n} p(x_i|E)}\right)\right]$

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.

Simplicity: any function that measures how straightforward an explanation is. Examples include parsimony, concision, and elegance. The appropriate choice of will generally depend on context.
Parsimony: a type of simplicity that reflects the number of elements, parameters, or principles an explanation requires.
Concision: a type of simplicity that measures how compact an explanation is, *e.g.*, by counting the number of words required to communicate it.

Box 2: Explanatory Values in Action



A common paradigm to tease apart explanatory values is disease diagnosis [22, 23, 24]. Participants are asked to explain a patient's symptoms by reference to different medical conditions. The figure above 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 the framework of Bayesian explanatory values proposed here, 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 implies a less simple relationships 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 (patients A and C fall 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, red also has higher co-explanation than yellow, since a patient with A's temperature under the yellow explanation can have a wider oxygenation range.

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 remains constant.

IV Explanation through the Lens of Co-explanation and Unification

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 a pattern of observations over and above how well it accounts for each independently. While this definition arises naturally from our Bayesian decomposition, its mathematical form matches that proposed by Ref. [25] as an operationalization of IBE and by Ref. [26] as an operationalization of explanatory considerations in Ref. [27].

Co-explanation is high when an explanation says some features of the data are predicted by others. For example, when economists observed that unemployment and inflation were inversely correlated, they proposed that the relationship was a general law—the Phillips curve [28]. Explanations that included this law had, as a consequence, co-explanatory value: inflation appeared to become predictable given knowledge of unemployment. Beyond economics, this value is particularly relevant in domains characterized by correlating common causes, such as diseases in medical diagnosis [29], legal cases [30], and social interactions [31]; see Box 2.

Psychological studies in these domains are often implicit tests of an individual's sensitivity to co-explanation as an explanatory value. Ref. [22] 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 a likely result (descriptiveness), but also whether their presentation preserved correlations between symptoms seen in the training set (co-explanation).[3]

That the mind links distinct experiences into coherent wholes is also at the heart of Gestalt psychology: "to apply the gestalt category means to find out which parts of nature belong as parts to functional wholes, to discover their position in these wholes, their degree of relative independence, and the articulation of larger wholes into sub-wholes" [32]. In vision, for example, it is argued that we perceive not the individual pieces of raw sense-data, but rather the explanation that links them together; the Kanizsa Triangle is compelling because the implicit shape we see co-explains the orientations of the missing wedges.

Co-explanation has the parallel theoretical value of unification, the expected co-explanation of the explanation conditional upon its truth. Unification says that the world is characterized not by coincidences, but patterns. It is a commonly value in the philosophy of science; for Ref. [33], a good scientific theory makes the manifestation of different phenomena dependent on each other, and a similar account can be found in Ref. [34], for whom good theories form a "systemtic" picture of the order of nature.

As with power, unification may be both virtue and vice. Even when a unifying theory is complicated, it does assert that the world itself is simple, because knowledge of some of its features allow you to predict others. Unifying theories, like powerful ones, are also more testable: one can look not only for unexpected events, but also patterns.[4] On the other hand, experiments show that unification may be perceived as a vice. Consider, for example, two explanations for "why Lois painted her nails in the shower" [35]: (a) 'she is afraid of spilling nail polish on her antique bathroom rug" or (b) "she is obsessive–compulsive",'. Explanation (b) correlates Lois' behavior with a many other (as yet) unobserved behaviors, and so is higher in unification (in that paper's terminology, has broader latent scope) than Explanation (a). However, subjects tended to prefer theories with lower unification (narrower latent scope).

[3] Crucially, co-explanation requires variation along different dimensions: if it is impossible, for example, for the data vary along a particular axes under a certain explanation, than knowledge of that aspect of the world tells us nothing new about the others because there is nothing left to explain. In this special case, the theory may have many virtues, such as high descriptiveness or power, but its co-explanation is at a minimum.

[4] More formally, the number of tests of a high power theory is linear in the number of features one looks at, but the number of tests of a unifying theory scale quadratically if one looks for pairwise correlations, or exponentially, when considering all combinations.

V Explanation through the Lens of Simplicity and Other Priors

Explanation and descriptiveness help us weigh explanations in a Bayesian fashion. Without simplicity, however, they are rarely good enough; as pointed out by thinkers such as Galileo, Newton, and Kant [36], one can often "improve" an explanation by adding more parameters, exceptions, and mechanisms, an observation linked to Occam's Razor. In Bayesian inference, simplicity enters into the prior. Accordingly, our intuitive notion of simplicity is captured by one or more theoretical values [37].

Ref. [23] finds evidence for this using an alien disease paradigm that puts descriptiveness and simplicity (here, **parsimony**) into conflict: participants' explanatory preferences are consistent with a value for descriptiveness plus a constant, *i.e.* data-independent, penalty for theory elaborateness (lack of parsimony). Further research has shown that such preferences take form early in childhood development and are robust across contexts [1, 2, 5].

Simplicity is complex, and our intuitive notion is an amalgam of many theoretical considerations. Table 3 illustrates the independent operation of two such considerations when explanations have a causal form. The upper-left theory is both more parsimonious (fewer hidden causes) and also more unified (providing a joint account of events) than the bottom-right theory. However, these effects can be decomposed. We therefore also have parsimonious-disjointed theories (where each visible aspect has a streamlined, but non-overlapping latent causal structure) and elaborate-unified theories (where everything is connected by a sprawling web of relations). This latter kind is reminiscent of conspiracy theories, where in one sense everything is very simple (all visible aspects are connected to each other), but that simplicity is achieved by postulating an elaborate web of hidden connections.

In statistics, simplicity is part of model selection, and there are many forms [21, 38]: the commonly-used Akaike Information Criterion [39] and Bayesian Information Criterion [40] are parsimony measures that count the number of parameters in a theory. In machine learning, regularization terms penalize non-zero parameter values to prevent overfitting to data. Meanwhile, the maximum entropy principle [41] understands simplicity as (negative) precision. There is general consensus both (1) simplicity is crucial to normative decision making, and (2) however the value is measured, it ought to enter additively in log-space, *i.e.*, as term in the prior shown in our Box 1 [42].

While statistical models are often concerned with prediction, the psychological value of simplicity goes well beyond this goal. Simple explanations can be easier to remember and work with, and may be preferred because we are limited beings with cognitive constraints. A simple explanation may be better because it requires fewer cognitive resources to apply or leads to fewer mistakes. Simple explanations are also easier to communicate and teach, and thus ease social coordination. Ssimplicity may even be seen as an aesthetic value, with simple explanations described as "elegant" and, on that basis, valuable in and of themselves; "mathematical beauty" plays a significant [43, 44], if controversial [45], role in physics.

Simplicity is not the only domain-general theoretical value, as there are many prior reasons to prefer one theory over another. Entire classes of explanations may be categorically better than others; for example, those that reveal causal mechanisms [46, 47, 48], explain new phenomena by analogy to familiar ones [49], or feature concrete mechanisms rather than abstract principles [29]. Within each class, humans have been shown to hold strong, contextually-informed priors for certain kinds of explanations. For example, we prefer explanations that involve diseases causing symptoms over the reverse [50] and find some causal connections immediately implausible [51]. These priors reflect our ability to flexibly apply background knowledge to new problems [52] and often take the form of intuitions and instincts—what Galileo referred to as "il lume naurale" or "the natural light" that guides our reason (1.80 of Ref. [53]).

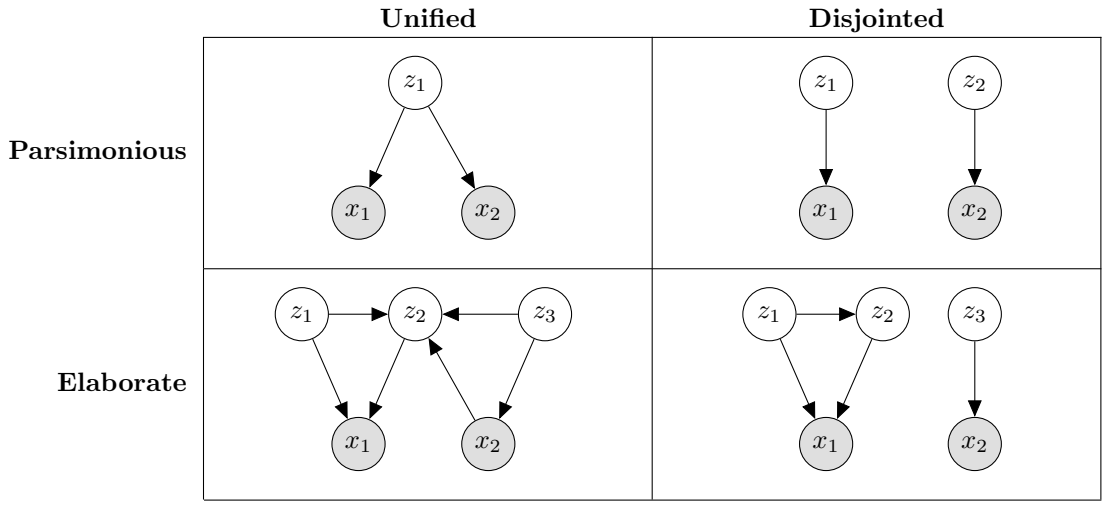


Table 1: In these diagrams that represent the causal structure of different explanations, shaded nodes represent observables, other nodes represented postulated latent causes, and arrows represent causal relationships. Simplicity is a rich concept that likely binds together multiple intuitions simultaneously. 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 an additive composite of the two.

VI When Values become Vices

Our approach yields a normative prescription where deviations from equal weighting lead to characteristic explanatory pathologies which canoperationalize what Ref. [54] calls "vice epistemology".

Consider the phenomenon of overgeneralization [55], *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).

What makes a weighting virtuous varies across contexts. For example, the appropriate amount of simplicity depends upon the domain [13], and there is evidence that people complexity-match [56], i.e., allow the perceived complexity of the explanandum to guide priors on the simplicity of the explanation. With that being said, recent empirical work has started to tie abnormal reasoning to common inferential biased that generalize across domains in a way that suggests the systemic miscalibration of values may be at fault.

For example, as noted by Ref. [57], those prone to paranormal thinking also show susceptibility to the conjunction fallacy. This can result from overvaluing co-explanation, because labeling Linda a feminist as well as a bank teller [58] provides a co-explanatory account of her political and social activities. Empirical work has also established strong individual differences in the tendency to believe conspiracies: those who believe one are more likely to believe others [59]. This trait is common in individuals with schizotypal disorders [60], which are linked in turn to a number of other explanatory abnormalities [61]. The finding that conspiracy-mindedness is a stable trait suggests that the its associated beliefs may be accumulated over time due, at least in part, to a systematic miscalibration in an individual's weighting of explanatory values.

Conspiracy theories are often both abnormally co-explanatory *and* descriptive [62]. They account for anomalous facts which are unlikely under the "official" explanation ("errant data" [63];

see, *e.g.*, Ref. [64] on Oaklahoma City bombing conspiracy theories), and show how seemingly arbitrary facts of ordinary life are correlated by hidden events [65]; Ref. [66] finds that a manipulation which induces subjects to see (illusory) correlations in neutral domains like stock returns also increases beliefs in conspiracy theories. 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 [67].

Valuing these features is not, in and of itself, a vice. What frequently goes wrong is the failure to balance these values against others, such as simplicity or contextual priors that urge trust in institutions and down-weight generalizations associated with racist, sexist, or antisemitic prejudice. On the surface, a conspiracy theory is quite 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 judgement is itself open to criticism; as noted by [68], some conspiracies are extremely compelling on normative grounds. Some even turn out to be true.

This latter point gets to the heart of what makes explanation so difficult. Striking a virtuous balance between so many considerations is itself a challenging cognitive problem, one that we solve partially by social circumspection. Failures at this level might help to explain anti-vaccination movements [69], COVID-19 conspiracies [70], the use of pseudoscience in extremist ideologies [71], and science denialism [72]. While these beliefs are in part formed and maintained by social processes in addition to epistemic ones, their core logic often appeals to many of the same explanatory imbalances as conspiracy theories do. These interact with individual-level predispositions, including what are usually taken to be pathologies of thought. One avenue for future research is how social processes may serve to maintain, accentuate, or exploit individual-level explanatory imbalances.[5]

VII Concluding remarks

Framing explanatory values as components of a Bayesian inference is a form of *rational analysis*, which seeks to understand mental states in terms of the computational goals they help agents achieve [82, 83]. Such an approach has been applied to a wide range of subjective states such as representativeness [84], suspicious coincidence [85], randomness [86], tip of the tongue [87], boredom and flow [88], mental effort [89], and curiosity [90, 91, 92]. Of these, explanation is most closely related to curiosity. If curiosity drives us to seek answers to salient questions [93] and to make sense of the world around us [94], then explanatory values are the subjective states that signal, often in compelling hedonic form [95], when good answers have been found.

A Bayesian framing naturally centers on an explanation's ability to predict observed data. Explanation is more than prediction, however, and other features are necessary to satisfy the many social, cognitive, and practical constraints that bear on the practice. A highly-predictive black box, for example, is not something that we can evaluate in terms of theoretical constructions such as parsimony or unification. Even when the black box is opened, what is inside may be so far from "virtuous" in the human sense that it scarcely counts as an explanation at all—even if it is intelligible in a literal sense. There is something more basic yet, of course: an explanation must be intelligible before we can ask about its value. This is part of the "explainability crisis" in machine learning [96] and is crucial to understanding, and thereby closing, the gap between human and artificial intelligence [97, 98]. While a rational analysis of explanatory values is an important first step, further work is needed to address the intelligibility problem.

All explanation occurs against a background of folk theories, world-views, and explanations that have come before [34]. This opinion suggests that it may be possible to enumerate "atomic" explanatory values, and that the history of explanation is largely the history of their relative emphasis. Given that explanations emerge in a social context, however, we might also expect new values—especially theoretical ones—to appear over time. This dynamical, contextual natural of

[5] Vices of overvaluation naturally co-exist with other biases, such as, for example, omission bias in the case of anti-vaccine movements [69].

Box 3: Three Stages of Explanation

	1. Generation	2. Selection	3. Evaluation
Observations:	fixed	fixed	variable
Explanations:	variable	fixed	fixed

Explanation decomposes into three stages: we generate explanations (or receive them from others), select among them, and re-evaluate them based on subsequent experience. Our piece has focused on how values influence selection, but they are equally important in generation and re-evaluation.

Values help us decide which experiences to seek out next. Co-explanation may lead me to look in places where I expect the data to be correlated under a favored hypothesis. Descriptiveness, conversely, may lead one to look for key counter-examples, in the style of Karl Popper's falsification [73].

Values also act at the generation stage. Ref. [74] describes scientific hypothesis formation as descriptiveness-driven, where explanations are updated in response to outliers: one produces an explanation, looks for places that it fails, and tries to update it in response. Co-explanation in the generation phase can also look like Peircean abudction [75]. The cycle, augmented to include the gathering of data, is used by both children [5] and adults [76].

A descriptiveness-driven cycle need not be virtuous: updating an explanation may increase its descriptiveness at the cost of theoretical values such as unification or simplicity. Kuhn's "paradigm shift" is driven in part by the decreasing simplicity of the standard paradigm: as anomalies arrive, more and more epicycles are required to explain them [77].

Rather than looking for outliers because we value descriptiveness, one may look for correlations because we value co-explanation. A person who subscribes to a "unifying" group stereotype, for example, may ask if people belong to the group in question when they show its characteristic behaviors. More virtuously, co-explanation can drive scientists to compare evidence across different domains.

Empirical values also direct attention at a more basic, cognitive level. Ref [78] find that descriptiveness draws the eye to outliers: attention to a group of pixels correlates with deviations from its predicted distribution. Co-explanation, conversely, draws attention to patterns that constitute gestalts [79].

Theoretical values, conversely, are crucial to how we go about generation, because we cannot consider every explanation. Parsimony and contextual considerations help us reject certain types of explanations out of hand [80], while unification makes the world itself easier to remember and describe. Theoretical values can sometimes be a vice, and we often fail to generate good explanations even when we have the ability to recognize them [81].

explanation are clarifies why explanations seem more valuable when they co-explain phenomena that, on the basis of previous understanding, were conceptually distant (see Box 3). Indeed, conceptual distance and co-explanation may be two sides of the same coin: what is conceptually distant may just be what, with our current explanations, we can not co-explain. Another important aspect of the dynamical side of explanation is the role of prediction out of sample, *i.e.*, re-evaluation when new information arrives. Because predicting the future is much more challenging than accounting for what is already known, doing so can be a powerful source of (empirical) value.

The importance of conceptual distance and the power of confirmation by unexpected data come together in "consilience". Consilience is an explanatory value introduced by [99] in the 19th Century to describe features of scientific explanations that, he argues, both are, and ought to be, prized by the community. "The Consilience of Inductions", Whewell writes, "takes place when an induction, obtained from one class of facts, coincides with an induction, obtained from another class... 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." Indeed, for Whewell, consilience carries simplicity and unification along for the ride: for consilient explanations, "all the additional suppositions tend to simplicity and harmony... the system becomes more coherent as it is further extended. The elements which we require for explaining a new class of facts are already contained in our system. Different members of the theory run together, and we have thus a constant convergence to unity."

Whewell is far from the only writer to draw attention to the general features of what makes for good explanation, and a significant part of social interaction involves debating and arguing for different explanations on the basis of the values they exhibit [100]. The implicit bargain for this paper is that such values may be amenable to an analysis in terms of basic, atomic units active in similar ways across a great variety of domains. Once identified, these units can provide a new way to understand how people make sense of the world.

Acknowledgements

We acknowledge the support of the John Templeton Foundation, and Jaan Tallinn via the Survival and Flourishing Fund.

Outstanding Questions

- How are the atomic elements of explanatory values perceived and combined by the mind? Why these elements, and not others? Where do they come from in evolutionary and cultural time?
- 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?
- 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 co-evolve 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 how 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 extremist ideologies) 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 is the relationship between moral and practical explanation?
- What makes an explanation intelligible "to us"?
- How can we enable machines to explain as well as predict? Can explanatory values help the task of bridging the gap between human and artificial intelligence?

References

[1] Tania Lombrozo. Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, 20(10):748–759, 2016.

[2] Caren M Walker, Elizabeth Bonawitz, and Tania Lombrozo. Effects of explaining on children's preference for simpler hypotheses. *Psychonomic bulletin & review*, 24(5):1538–1547, 2017.

[3] Ala Samarapungavan. Children's judgments in theory choice tasks: Scientific rationality in childhood. *Cognition*, 45(1):1–32, 1992.

[4] Brandy N Frazier, Susan A Gelman, and Henry M Wellman. Young children prefer and remember satisfying explanations. *Journal of Cognition and Development*, 17(5):718–736, 2016.

[5] Elizabeth Baraff Bonawitz and Tania Lombrozo. Occam's rattle: Children's use of simplicity and probability to constrain inference. *Developmental psychology*, 48(4):1156, 2012.

[6] Gilbert H Harman. The inference to the best explanation. *The philosophical review*, 74(1):88–95, 1965.

[7] Thomas L Griffiths, Nick Chater, Charles Kemp, Amy Perfors, and Joshua B Tenenbaum. Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in cognitive sciences*, 14(8):357–364, 2010.

[8] Adam N Sanborn and Nick Chater. Bayesian brains without probabilities. *Trends in cognitive sciences*, 20(12):883–893, 2016.

[9] Jonah N Schupbach. Inference to the best explanation, cleaned up and made respectable. *Best explanations: New essays on inference to the best explanation*, 2016.

[10] Igor Douven and Sylvia Wenmackers. Inference to the best explanation versus bayes's rule in a social setting. *The British Journal for the Philosophy of Science*, 68(2):535–570, 2017.

[11] Igor Douven. Inference to the best explanation: What is it? and why should we care. *Best explanations: New essays on inference to the best explanation*, pages 4–22, 2017.

[12] Nevin Climenhaga. Inference to the best explanation made incoherent. *The Journal of Philosophy*, 114(5):251–273, 2017.

[13] Michael Huemer. When is parsimony a virtue? *The Philosophical Quarterly*, 59(235):216–236, 2009.

[14] Peter Lipton. *Inference to the best explanation*. Taylor & Francis, 2004.

[15] Jonathan Weisberg. Locating ibe in the bayesian framework. *Synthese*, 167(1):125–143, 2009.

[16] Leah Henderson. Bayesianism and inference to the best explanation. *The British Journal for the Philosophy of Science*, 65(4):687–715, 2014.

[17] Bas C Van Fraassen. Laws and symmetry. 1989.

[18] Stathis Psillos. Inference to the best explanation and bayesianism. In *Induction and deduction in the sciences*, pages 83–91. Springer, 2004.

[19] Milan Studenỳ and Jirina Vejnarová. The multiinformation function as a tool for measuring stochastic dependence. In *Learning in graphical models*, pages 261–297. Springer, 1998.

[20] Benjamin Enke and Florian Zimmermann. Correlation neglect in belief formation. *The Review of Economic Studies*, 86(1):313–332, 2019.

[21] Edwin T Jaynes. Inference, method, and decision: Towards a bayesian philosophy of science, 1979.

[22] Douglas L Medin, Mark W Altom, Stephen M Edelson, and Deborah Freko. Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8(1):37, 1982.

[23] Tania Lombrozo. Simplicity and probability in causal explanation. *Cognitive psychology*, 55(3):232–257, 2007.

[24] Samuel Johnson, Angie Johnston, Amy Toig, and Frank Keil. Explanatory scope informs causal strength inferences. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36, 2014.

[25] Wayne C Myrvold. A bayesian account of the virtue of unification. *Philosophy of Science*, 70(2):399–423, 2003.

[26] Jiji Zhang and Kun Zhang. Likelihood and consilience: on Forster's counterexamples to the likelihood theory of evidence. *Philosophy of Science*, 82(5):930–940, 2015.

[27] Malcolm R Forster. Unification, explanation, and the composition of causes in newtonian mechanics. *Studies In History and Philosophy of Science Part A*, 19(1):55–101, 1988.

[28] Alban W Phillips. The relation between unemployment and the rate of change of money wage rates in the united kingdom, 1861–1957 1. *economica*, 25(100):283–299, 1958.

[29] Stefan Dragulinescu. Inference to the best explanation and mechanisms in medicine. *Theoretical medicine and bioethics*, 37(3):211–232, 2016.

[30] Amalia Amaya. Inference to the best legal explanation. In *Legal Evidence and Proof*, pages 149–174. Routledge, 2016.

[31] Max Kleiman-Weiner, Mark K Ho, Joseph L Austerweil, Michael L Littman, and Joshua B Tenenbaum. Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction. In *CogSci*, 2016.

[32] Kurt Koffka. Gestalt. In *Encyclopaedia of the Social Sciences*. 1931.

[33] Michael Friedman. Explanation and scientific understanding. *The Journal of Philosophy*, 71(1):5–19, 1974.

[34] Philip Kitcher. Explanatory unification and the causal structure of the world. 1989.

[35] Sangeet S Khemlani, Abigail B Sussman, and Daniel M Oppenheimer. Harry potter and the sorcerer's scope: latent scope biases in explanatory reasoning. *Memory & cognition*, 39(3):527–535, 2011.

[36] Alan Baker. Simplicity. 2004.

[37] Irving John Good. Corroboration, explanation, evolving probability, simplicity and a sharpened razor. *The British Journal for the Philosophy of Science*, 19(2):123–143, 1968.

[38] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[39] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[40] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[41] Edwin T Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

[42] IJ Good. Explicativity: a mathematical theory of explanation with statistical applications. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 354(1678):303–330, 1977.

[43] Brian Greene. The elegant universe: Superstrings, hidden dimensions, and the quest for the ultimate theory, 2000.

[44] Elaine Scarry. *On beauty and being just*. Princeton University Press, Princeton, NJ, USA, 2013.

[45] Sabine Hossenfelder. *Lost in math: how beauty leads physics astray*. Basic Books, 2018.

[46] Wesley C Salmon. *Causality and explanation*. Oxford University Press, 1998.

[47] Jeffrey C Zemla, Steven Sloman, Christos Bechlivanidis, and David A Lagnado. Evaluating everyday explanations. *Psychonomic bulletin & Review*, 24(5):1488–1500, 2017.

[48] Jeffrey Zemla, Steven A Sloman, Christos Bechlivanidis, and David Lagnado. Not so simple! mechanisms increase preference for complex explanations. 2020.

[49] Paul R Thagard. The best explanation: Criteria for theory choice. *The journal of philosophy*, 75(2):76–92, 1978.

[50] Joshua B Tenenbaum, Thomas L Griffiths, and Sourabh Niyogi. Intuitive theories as grammars for causal inference. *Causal learning: Psychology, philosophy, and computation*, pages 301–322, 2007.

[51] Saiwing Yeung and Thomas L Griffiths. Identifying expectations about the strength of causal relationships. *Cognitive psychology*, 76:1–29, 2015.

[52] John D Norton. A material theory of induction. *Philosophy of Science*, 70(4):647–670, 2003.

[53] Charles Sanders Peirce. *Collected papers of charles sanders peirce*, volume 2. Harvard University Press, 1960.

[54] Quassim Cassam. Vice epistemology. *The Monist*, 99(2):159–180, 2016.

[55] Joseph Jay Williams, Tania Lombrozo, and Bob Rehder. The hazards of explanation: Overgeneralization in the face of exceptions. *Journal of Experimental Psychology: General*, 142(4):1006, 2013.

[56] Jonathan B. Lim and Daniel M. Oppenheimer. Explanatory preferences for complexity matching. *PLOS ONE*, 15(4):1–19, 04 2020.

[57] Robert Brotherton and Christopher C French. Belief in conspiracy theories and susceptibility to the conjunction fallacy. *Applied Cognitive Psychology*, 28(2):238–248, 2014.

[58] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131, 1974.

[59] Martin Bruder, Peter Haffke, Nick Neave, Nina Nouripanah, and Roland Imhoff. Measuring individual differences in generic beliefs in conspiracy theories across cultures: Conspiracy mentality questionnaire. *Frontiers in psychology*, 4:225, 2013.

[60] Hannah Darwin, Nick Neave, and Joni Holmes. Belief in conspiracy theories. the role of paranormal belief, paranoid ideation and schizotypy. *Personality and Individual Differences*, 50(8):1289–1293, 2011.

[61] Benjamin F McLean, Julie K Mattiske, and Ryan P Balzan. Association of the jumping to conclusions and evidence integration biases with delusions in psychosis: a detailed meta-analysis. *Schizophrenia Bulletin*, 43(2):344–354, 2017.

[62] Joseph A Vitriol and Jessecae K Marsh. The illusion of explanatory depth and endorsement of conspiracy beliefs. *European Journal of Social Psychology*, 48(7):955–969, 2018.

[63] Brian L. Keeley. Of conspiracy theories. *The Journal of Philosophy*, 96(3):109–126, 1999.

[64] David Coady. Conspiracy theories: the philosophical debate. In The editor, editor, *Conspiracy Theories: the Philosophical Debate*, volume 4 of *5*, chapter 8, pages 201–213. The name of the publisher, The address of the publisher, 3 edition, 7 1993. An optional note.

[65] Timothy Tangherlini. Toward a generative model of legend: Pizzas, bridges, vaccines, and witches. *Humanities*, 7(1):1, Dec 2017.

[66] Jennifer A Whitson and Adam D Galinsky. Lacking control increases illusory pattern perception. *science*, 322(5898):115–117, 2008.

[67] Karen M Douglas, Robbie M Sutton, and Aleksandra Cichocka. The psychology of conspiracy theories. *Current directions in psychological science*, 26(6):538–542, 2017.

[68] Matthew RX Dentith. When inferring to a conspiracy might be the best explanation. *Social Epistemology*, 30(5-6):572–591, 2016.

[69] Helena Miton and Hugo Mercier. Cognitive obstacles to pro-vaccination beliefs. *Trends in Cognitive Sciences*, 19(11):633–636, 2015.

[70] Shadi Shahsavari, Pavan Holur, Timothy R. Tangherlini, and Vwani Roychowdhury. Conspiracy in the time of corona: Automatic detection of covid-19 conspiracy theories in social media and the news, 2020.

[71] Rachel O'Neill. *Seduction: Men, masculinity and mediated intimacy*. John Wiley & Sons, 2018.

[72] Bastiaan T Rutjens, Steven J Heine, Robbie M Sutton, and Frenk van Harreveld. Attitudes towards science. In *Advances in Experimental Social Psychology*, volume 57, pages 125–165. Elsevier, 2018.

[73] Karl Popper. *The logic of scientific discovery*. Routledge, 2005.

[74] Andrew Gelman and Cosma Rohilla Shalizi. Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38, 2013.

[75] William HB McAuliffe. How did abduction get confused with inference to the best explanation? *Transactions of the Charles S. Peirce Society: A Quarterly Journal in American Philosophy*, 51(3):300–319, 2015.

[76] Neil R Bramley, Peter Dayan, Thomas L Griffiths, and David A Lagnado. Formalizing neurath's ship: Approximate algorithms for online causal learning. *Psychological review*, 124(3):301, 2017.

[77] Thomas S Kuhn. *The structure of scientific revolutions*. University of Chicago press, 2012.

[78] Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision research*, 49(10):1295–1306, 2009.

[79] Agnes Desolneux, Lionel Moisan, and Jean-Michel Morel. *From gestalt theory to image analysis: a probabilistic approach*, volume 34. Springer Science & Business Media, 2007.

[80] Alison Gopnik, Shaun O'Grady, Christopher G Lucas, Thomas L Griffiths, Adrienne Wente, Sophie Bridgers, Rosie Aboody, Hoki Fung, and Ronald E Dahl. Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National Academy of Sciences*, 114(30):7892–7899, 2017.

[81] Anselm Rothe, Brenden M Lake, and Todd M Gureckis. Do people ask good questions? *Computational Brain & Behavior*, 1(1):69–89, 2018.

[82] Falk Lieder and Thomas L Griffiths. Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, pages 1–85, 2019.

[83] Nick Chater and Mike Oaksford. Ten years of the rational analysis of cognition. *Trends in cognitive sciences*, 3(2):57–65, 1999.

[84] Joshua B Tenenbaum, Thomas L Griffiths, et al. The rational basis of representativeness. In *Proceedings of the 23rd annual conference of the Cognitive Science Society*, pages 1036–1041. Citeseer, 2001.

[85] Thomas L Griffiths and Joshua B Tenenbaum. From mere coincidences to meaningful discoveries. *Cognition*, 103(2):180–226, 2007.

[86] Thomas L Griffiths, Dylan Daniels, Joseph L Austerweil, and Joshua B Tenenbaum. Subjective randomness as statistical inference. *Cognitive psychology*, 103:85–109, 2018.

[87] Alan S Brown. A review of the tip-of-the-tongue experience. *Psychological bulletin*, 109(2):204, 1991.

[88] Nick Chater, George Loewenstein, and Zachary Wojtowicz. Boredom and flow: An opportunity cost theory of attention-directing motivational states. *Available at SSRN 3339123*, 2019.

[89] Robert Kurzban, Angela Duckworth, Joseph W Kable, and Justus Myers. An opportunity cost model of subjective effort and task performance. *Behavioral and brain sciences*, 36(6):661–679, 2013.

[90] George Loewenstein. The psychology of curiosity: A review and reinterpretation. *Psychological bulletin*, 116(1):75, 1994.

[91] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 16–17, 2017.

[92] George Loewenstein and Zachary Wojtowicz. Curiosity and the economics of attention. *Unpublished Manuscript*, 2020.

[93] Russell Golman and George Loewenstein. Curiosity, information gaps, and the utility of knowledge. *Information Gaps, and the Utility of Knowledge (April 16, 2015)*, 2015.

[94] Nick Chater and George Loewenstein. The under-appreciated drive for sense-making. *Journal of Economic Behavior & Organization*, 126:137–154, 2016.

[95] Alison Gopnik. Explanation as orgasm. *Minds and machines*, 8(1):101–118, 1998.

[96] Andrew Selbst and Julia Powles. "meaningful information" and the right to explanation. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 48–48, New York, NY, USA, 23–24 Feb 2018. PMLR.

[97] Michael Pacer, Joseph Williams, Xi Chen, Tania Lombrozo, and Thomas Griffiths. Evaluating computational models of explanation using human judgments. *arXiv e-prints*, 2013. 1309.6855.

[98] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv preprint arXiv:1904.12584*, 2019.

[99] William Whewell. *The Philosophy of the Inductive Sciences: Founded Upon Their History*, volume 2. John W. Parker, West Strand, London, United Kingdom, 2nd edition, 1847. Page 65.

[100] H. Mercier and D. Sperber. *The Enigma of Reason*. Harvard University Press, Cambridge, MA, USA, 2017.