Inside Ockham's Razor: A Mechanism Driving Preferences for Simpler Explanations

Thalia H. Vrantsidis[a], Tania Lombrozo[b]

[a]Department of Psychology, Mississippi State University

[b]Department of Psychology, Princeton University

This is a preprint of a manuscript accepted at the journal Memory and Cognition.

Author Note

Correspondence concerning this article should be addressed to Thalia H. Vrantsidis, Department of Psychology, Mississippi State University, Rice Hall, Mississippi State, MS, 39762, USA. Email: tvrantsidis@psychology.msstate.edu. ORCID: 0000-0003-0766-9041

Abstract

People often prefer simpler explanations, defined as those that posit the presence of fewer causes (e.g., positing the presence of a single cause, Cause A, rather than two causes, Causes B and C, to explain observed effects). Here we test one hypothesis about the mechanisms underlying this preference: that people tend to reason as if they are using "agnostic" explanations, which remain neutral about the presence/absence of additional causes (e.g., comparing "A" vs. "B and C", while remaining neutral about the status of B and C when considering "A", or of A when considering "B and C"), even in cases where "atheist" explanations, which specify the absence of additional causes (e.g., "A *and not B or C*" vs. "B and C *and not A*"), are more appropriate. Three studies with US-based samples (total N = 982) tested this idea by using scenarios for which agnostic and atheist strategies produce diverging simplicity/complexity preferences, and asking participants to compare explanations provided in atheist form. Results suggest that people tend to ignore absent causes, thus overgeneralizing agnostic strategies, which can produce preferences for simpler explanations even when the complex explanation is objectively more probable. However, these unwarranted preferences were reduced by manipulations that encouraged participants to consider absent causes: making absences necessary to produce the effects (Study 2), or describing absences as causes that produce alternative effects (Study 3). These results shed light on the mechanisms driving preferences for simpler explanations, and on when these mechanisms are likely to lead people astray.

Keywords: Simplicity, Complexity, Explanation, Explanatory Virtues, Probability, Ockham's Razor

Introduction

Consider the following question. There is an antique clock in a museum, and it has two timekeeping problems: it runs slow and it skips hours. These problems could be due to misplaced spring-plugs (Cause A), which would cause both problems, or to a combination of worn gears (Cause B), which would cause it to run slow, and a winding malfunction (Cause C), which would cause it to skip hours. What is the most likely explanation for the clock's timekeeping problems?

In such cases, existing work suggests that both children and adults prefer simpler explanations, i.e., those that posit the presence of fewer unexplained causes (e.g., Bonawitz & Lombrozo, 2012; Johnson et al., 2019; Lombrozo, 2007; Pacer & Lombrozo, 2017; Read & Marcus-Newhall, 1993). In the current example, that would mean explaining the clock's problems with a single cause (A: misplaced spring-plugs) rather than two causes (B & C: worn gears and a winding malfunction). This simplicity preference can be viewed as a form of Ockham's razor, the well-known stricture to favor simpler explanations (all else being equal). Yet, while simplicity preferences in these cases appear to be quite robust, the mechanisms that generate these preferences are not well-understood. Thus, the present paper aims to shed light on one aspect of this mechanism: how people represent and reason over the explanations being compared, and, in particular, the causes that are absent in those explanations.

Previously Observed Preferences for Simpler Explanations

To examine the mechanisms underlying simplicity preferences, the current work focused on cases like the clock example, which have been used frequently in prior work (Bonawitz & Lombrozo, 2012; Johnson et al., 2019; Lombrozo, 2007; Pacer & Lombrozo, 2017; Read & Marcus-Newhall, 1993; Zemla et al., 2023). These cases are characterized by the following features. Participants are given a causal structure like that in the clock scenario, where one cause

(e.g., misplaced-spring plugs) can produce multiple effects, or multiple causes (e.g., worn gears and a winding malfunction) each produce one of these effects. Participants are then asked to identify the best, most satisfying, or most probable explanation for a given set of effects (e.g., the clock's timekeeping issues), where the potential explanations vary in terms their simplicity (in this example, whether they posit one vs. two causes). More precisely, simpler explanations are defined as those that posit the presence of fewer unexplained causes (i.e., instances of a cause being present, where this is not explained by positing the presence of another cause). This definition of simplicity is well-suited to examining cases like the clock example, where the task is to explain a specific event by positing the presence of specific instances of causes (e.g., explaining a specific clock's issues by positing that it has misplaced spring-plugs), rather than positing the existence of novel types of causes (as in many scientific explanations). When examining cases with these characteristics, previous work has shown a fairly robust preference for simpler explanations. This preference is especially robust in cases where the simpler explanation is indeed more probable (given elicited or provided probability information, or reasonable assumptions about the probabilities; e.g., Johnson et al., 2019; Lombrozo, 2007; Pacer & Lombrozo, 2017; Read & Marcus-Newhall, 1993; Vrantsidis & Lombrozo, 2022). Moreover, the preference for simpler explanations is sometimes overgeneralized to cases for which the *complex* explanation is instead more probable (Bonawitz & Lombrozo, 2012; Lombrozo, 2007; Pacer & Lombrozo, 2017, 2017; Shimojo et al., 2020). Notably, there are also cases for which complex explanations are actually preferred, though these cases typically use tasks and stimuli that differ from the clock example in various ways (e.g., Lim & Oppenheimer, 2020; Liquin & Lombrozo, 2022; Marsh et al., 2022; Zemla et al., 2017).

Mechanisms Underlying Simplicity Preferences: Using Agnostic vs. Atheist Explanations

In the present work, we used cases like the clock scenario, for which simplicity preferences are well-established, to examine an unexplored aspect of the mechanisms driving simplicity preferences. Specifically, we examine how people represent and reason over the explanations being compared, and, in particular, the causes that are absent in those explanations. For example, when comparing the simple and complex explanation in the clock scenario, do people consider only those causes posited to be *present* by each explanation (thus comparing "A", misplaced spring-plugs, to "B & C", worn gears and a winding malfunction), ignoring or remaining neutral about the presence of other potential causes (e.g., ignoring B and C when considering the simple explanation)? Or do people also consider the *absence* of other causes (thus perhaps comparing "A & not B & not C", i.e., the clock having misplaced spring-plugs and *not* worn gears or a winding malfunction, to "B & C & not A", i.e., the clock having worn gears and a winding malfunction, but *not* misplaced spring-plugs)? Following Sober (2006), we will refer to explanations as "agnostic" if they only stipulate causes that are present, while remaining neutral about the presence versus absence of additional causes, and we will refer to explanations as "atheist" if they stipulate the absence of additional causes.[1] (This is analogous to how agnostics remain neutral on the existence of God, while atheists deny God's existence.) Moreover, we define an "agnostic" explanation-evaluation strategy as one in which people act as if they are reasoning over agnostic explanations: either because they do not represent absent

[1] The definitions of atheist/agnostic explanations merely refer to whether the explanations include only present causes, or also absent causes, and thus do not specify *which* of all possible present or absent causes are included. However, throughout we assume that some process of variable selection allows people to focus on causes that are at least potentially relevant (e.g., a volcano on Mars is unlikely to be considered as a cause of a clock running slow on Earth; Henne et al., 2017; Hesslow, 1988; Kinney & Lombrozo, 2022). Studies 2 and 3 focus on how this selection process might affect whether absent causes are considered, and, in the General Discussion, we address how a focus on present rather than absent causes might also contribute to the variable selection process.

causes,[2] because they represent those causes but remain neutral about their presence/absence, or because they do not consider absences in their reasoning process. In contrast, we define "atheist" strategies as ones in which people act as if they are reasoning over atheist explanations, which presumably involves both representing absent causes and accounting for these causes in one's reasoning. We thus ask whether previously observed simplicity preferences stem from using *agnostic* vs. *atheist* strategies for evaluating explanations.

To see how using atheist vs. agnostic strategies might affect explanation preferences, we will use Bayesian inference as a framework to formalize the consequences of reasoning using these two types of explanations. To start, we assume that when comparing explanations, people are trying to compare the explanations' posterior probabilities, P(Explanation | Effects): the probability of the explanation being true (i.e., all causes having their hypothesized states), given the observed effects that are to be explained. Although this will not necessarily fully capture explanation judgments (e.g., Pacer et al., 2013), it offers a useful benchmark from which we can evaluate departures. In cases like the clock example, applying this Bayesian approach would mean comparing P(Cause A present | Effects) vs. P(Causes B and C present | Effects) if using agnostic explanations, or comparing P(Cause A present and not B and not C | Effects) vs. P(Causes B and C present and not A | Effects) if using atheist explanations. According to Bayes'

2 One might wonder whether it is plausible for people to not represent the absent causes at all in the examples we consider here. For example, suppose someone was considering the complex explanation "Causes B and C", and completely failed to represent Cause A. That is, this person is excluding Cause A from their causal model (as shown in the diagram in Table 1; rather than including it in their model but remaining neutral about its status). This person is at least implicitly treating the status of Cause A as causally irrelevant. One might wonder whether this is even plausible: presumably omitting Cause A from their model would then lead the person to say that changing the status of Cause A would never have any effect on the observed effects (even if Causes B and C were absent), which seems unreasonable. Instead, we suggest that if people are failing to represent the absent cause, Cause A, they are doing so temporarily, while they are evaluating the complex explanation. If they were then asked to evaluate the simple explanation, where only Cause A is specified as present, they would presumably rebuild a new mental model in which Cause A was included and viewed as causally relevant, but Causes B and C would be ignored. The current work thus views this as one possible version of an agnostic strategy, but does not try to distinguish it from others (e.g., representing absent causes but not representing their status, or representing absent causes but ignoring them in one's reasoning process).

rule, the posterior probabilities for competing explanations can be compared by using two other pieces of information: the explanations' *prior probabilities*, P(Explanation), and *likelihoods*, P(Effects | Explanation), according to the following formula:

$$\frac{P(\,Explanation_1 \mid Effects\,)}{P(\,Explanation_2 \mid Effects\,)} = \frac{P(\,Effects \mid Explanation_1\,)}{P(\,Effects \mid Explanation_2\,)} * \frac{P(\,Explanation_1\,)}{P(\,Explanation_2\,)} \quad (1)$$

The prior probabilities here reflect the chance of the explanation being true (i.e., the joint probability of the causes taking their hypothesized values), *not* conditioned on the observed effects. For example, for the explanation "Causes A and B present," this would be the chance of Causes A and B being present, in general (e.g., of a clock having worn gears and a winding malfunction in general, without knowing anything about whether it has particular time keeping issues). The explanations' likelihoods reflect the chance of the observed effects occurring (i.e., their joint probability) if the explanation were true (i.e., assuming the causes took on their hypothesized values): e.g. the chance of a clock having those particular timekeeping issues (running slow and skipping hours), assuming a given explanation is true (say, that it had worn gears and a winding malfunction). Using this framework, we can then compare the consequences of reasoning over atheist vs. agnostic explanations in cases like the clock example. By working through the math, we can see that these two strategies will often produce similar consequences: in particular, there are many assumptions about priors and likelihoods for which both atheist and agnostic explanations converge in producing preferences for simpler explanations. As one example, suppose causes are equally rare (say, each has a 20% chance of being present), statistically independent, and guaranteed to produce their effects. Working through the math for these conditions (see Table 1), we see that simplicity preferences arise for both atheist and agnostic explanations – that is, the simpler explanation has a higher posterior probability in both

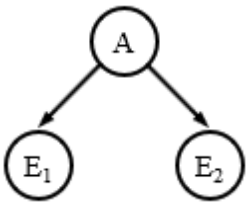
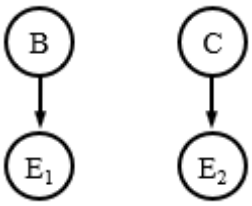
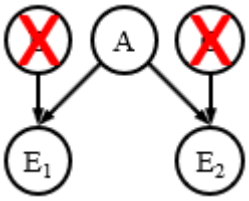
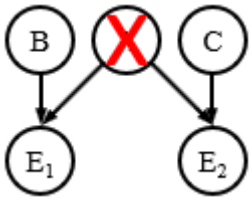
cases, which is driven by its higher prior probability. Intuitively, this is because when causes are present only rarely, it is generally more likely for one cause to be present, rather than two.

In contrast, there are also conditions in which using atheist vs. agnostic explanations can lead to diverging relative posteriors in cases like the clock example. For instance, suppose we modify our previous example so that all causes are instead equally *common* (say, each has an 80% chance of being present, rather than a 20% chance). Working through the math (see Table 1) we see that treating explanations as agnostic again produces a preference for the simpler explanation, but now treating explanations as atheist instead produces a preference for the more *complex* explanation. Intuitively, this is because the absence of a cause is now rare, and the simple atheist explanation ("A & not B & not C") contains two absences (which should make it especially unlikely), while the complex atheist explanation ("B & C & not A") only contains one absence. In contrast, when using agnostic explanations, the probability of these absences does not factor into the computations, so one cause being present continues to be more likely than two causes being present. Performing similar computations for other cases shows that, while agnostic and atheist strategies will often agree, there are a range of conditions under which they diverge, such that an agnostic strategy predicts a simplicity preference while an atheist strategy predicts a complexity preference (see Figure 1).

Table 1

Properties of the Simple and Complex Explanations in the Current Scenarios under Atheist and Agnostic Interpretations.

Explanation Type	Explanation	
	Simple	Complex

Agnostic	 "Cause A"	 "Causes B & C"
Atheist	 "Cause A & not B & not C"	 "Causes B & C & not A"

If each cause is present 20% of the time:

	Explanations' prior probabilities	
Agnostic	P(A) = 20%	P(B) * P(C) = 20% * 20% = 4%
Atheist	P(A) * P(not B) * P(not C) = 20% * 80% * 80% = 12.8%	P(not A) * P(B) * P(C) = 20% * 20% * 80% = 3.2%

	Ratio of posterior probabilities
Agnostic	20% : 4% = 5:1
Atheist	12.8% : 3.2% = 4:1

	Predicted explanatory preference
Agnostic	Simplicity preference
Atheist	Simplicity preference

If each cause is present 80% of the time:

	Explanations' prior probabilities	
Agnostic	P(A) = 80%	P(B) * P(C) = 80% * 80% = 64%
Atheist	P(A) * P(not B) * P(not C) = 80% * 20% * 20% = 3.2%	P(not A) * P(B) * P(C) = 80% * 80% * 20% = 12.8%

	Ratio of posterior probabilities
Agnostic	80% : 64% = 5:4
Atheist	3.2% : 12.8% = 1:4

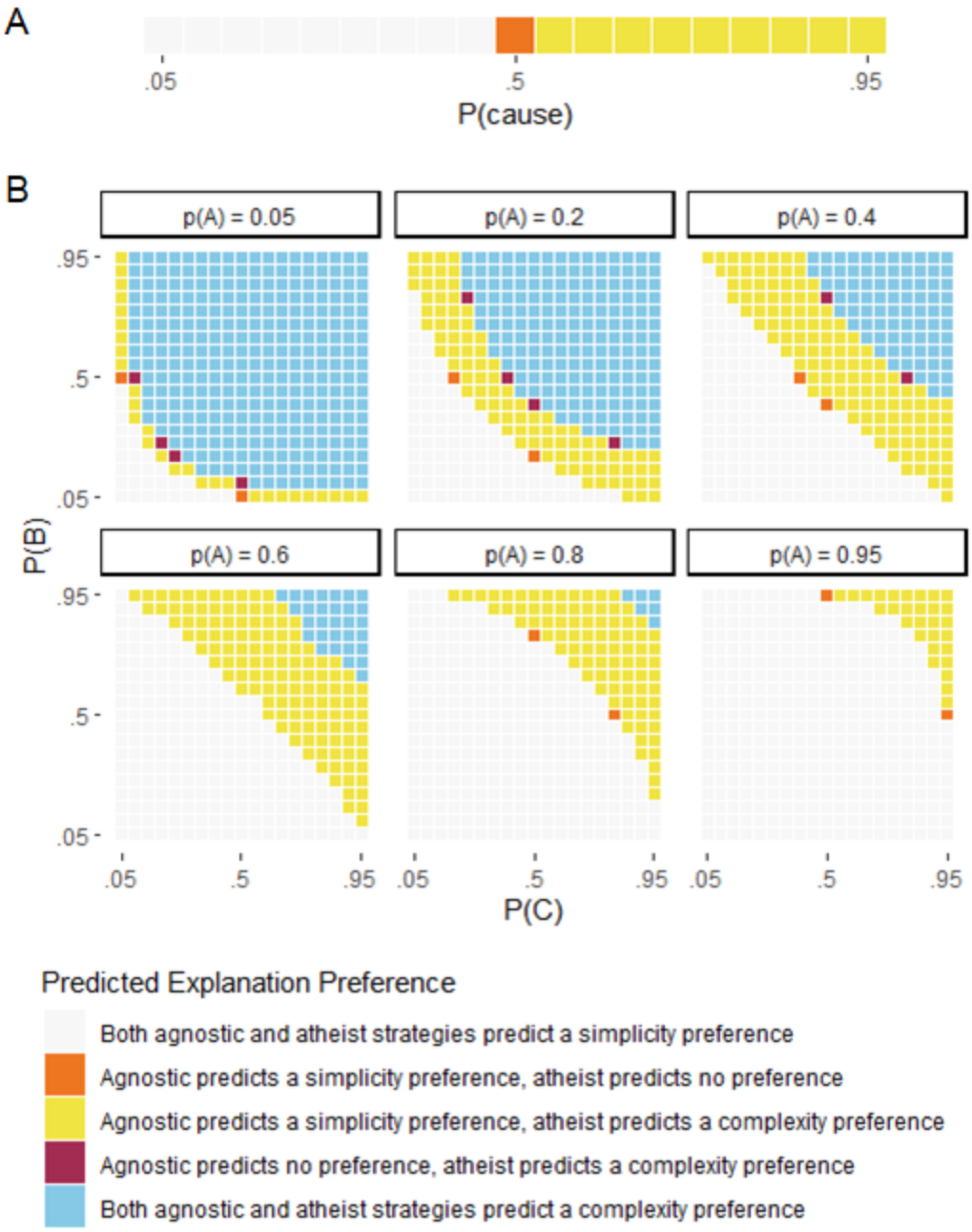
	Predicted explanatory preference
Agnostic	Complexity preference

Atheist	Simplicity preference

Note. Prior probabilities (i.e., explanations' probabilities, prior to considering the observed effects) are computed assuming that the probabilities of the three causes are statistically independent, so that, e.g., P(X & Y) = P(X)*P(Y). The ratio of the explanations' posterior probabilities (i.e., their probability, after considering the observed effects) is computed simply as the ratio of their prior probabilities, based on the assumption that causes are guaranteed to produce their effects, so that both the simple and complex explanations have Bayesian likelihoods of 100%, i.e. a 100% chance of producing the observed effects if the explanations are true. Equation 1 can therefore be simplified by inputting 1/1 as the ratio of the likelihoods, or, equivalently, removing the likelihoods from the equation. Predicted explanatory preferences are based on which explanation has a higher posterior probability. The diagram of the "agnostic" causal model in Table 1 reflects one of multiple possible ways in which an agnostic strategy could be implemented as a causal model (i.e., including present causes in the model and specifying them as present, while not including other causes). Alternatives include incorporating the other causes in the model without specifying their status as present or absent, or representing the absent causes but using a reasoning strategy that fails to account for them in computing posteriors.

Figure 1

Predictions of Agnostic and Atheist Strategies in terms of Simplicity/Complexity Preferences



Note. The regions in yellow, orange and purple show cases for which the two strategies diverge. In (A), predictions are based on the following assumptions: each of the three causes (Causes A, B, and C) are present with equal probability (where this probability varies, as shown on the x-axis), and the causes are statistically independent and guaranteed to produce their effects. In (B), this first assumption is relaxed so that the probability of each cause being present is allowed to

vary independently. The probability of Cause A being present varies across sub-plots, the probability of Cause B being present varies along the y-axis, and the probability of Cause C being present varies along the x-axis. Displayed probabilities range from .05 to .95, rather than 0 to 1, since, when comparing different explanations, it is unlikely that one would consider causes known to always be present or always be absent.

In the current work, we examine whether people tend to rely on atheist or agnostic strategies. In particular, we investigate the hypothesis that simplicity preferences often arise from an agnostic strategy for evaluating explanations, and, moreover, that people tend to overgeneralize this agnostic strategy to situations for which an atheist strategy is more appropriate – for example, because one is specifically asked about atheist explanations. To test which strategies people use, we examine cases where the predictions of agnostic and atheist strategies diverge (e.g., when all causes have an 80% chance of being present), such that agnostic strategies predict simplicity preferences while atheist strategies predict complexity preferences. Moreover, to test whether agnostic strategies are *overgeneralized*, we ask participants to compare explanations specified in atheist form (i.e., comparing "A & not B & not C" to "B & C & not A"), so that the most appropriate (or at least literal) interpretation of the question requires considering atheist explanations, and the absences specified within them.

Characterizing explanatory preferences under these conditions is valuable for a number of reasons. First, cases where agnostic and atheist strategies diverge offer a unique opportunity to uncover the mechanisms that drive simplicity preferences more generally. Observing simplicity preferences in such cases would offer strong evidence that explanation evaluation often proceeds using agnostic strategies, even when atheist strategies would be more appropriate. In contrast,

observing complexity preferences would indicate a role for atheist strategies in people's reasoning. Second, identifying which strategy is in use has implications for predicting and intervening on people's explanation preferences. Knowing that people overgeneralize agnostic strategies can be used to anticipate errors and biases in people's judgments, including biases towards "oversimplification": i.e., being biased towards simpler explanations, such that they are favored beyond what is mathematically justified given the available information (e.g., preferring a simpler explanation, or even showing no preference between explanations, when the provided information makes the complex explanation more probable). Anticipating these errors, and understanding their basis, can in turn potentially guide interventions to help people avoid oversimplification. In contrast, finding that people appropriately use atheist strategies in these cases would help demarcate boundary conditions on previously observed simplicity preferences, offering a basis for predicting when people might instead show complexity preferences. In sum, this work can help to shed light on the mechanisms underlying people's preferences for simpler (and perhaps oversimplified) explanations, as well as the conditions in which these preferences occur.

Challenges in Identifying Overgeneralization of Agnostic Strategies and Distinguishing between Agnostic vs. Atheist Strategies based on Previous Work

A number of previous studies have investigated simplicity preferences in cases like the clock example. In the populations tested, studies have typically found preferences for simpler explanations (see references in Table 2), and have additionally identified factors that attenuate (but rarely reverse) this preference (Johnson et al., 2019; Lombrozo, 2007; Pacer & Lombrozo, 2017; Shimojo et al., 2020; Vrantsidis & Lombrozo, 2022; Zemla et al., 2023). However, there are two main challenges with addressing the present set of questions using previous work. First,

in most of this work, it is not clear whether atheist or agnostic strategies are more appropriate, and thus whether one of these strategies is being overgeneralized. And, second, most existing work cannot clearly distinguish which of these strategies is being used at all. We discuss these two challenges in turn (see also Table 2).

First, our ability to know whether atheist or agnostic strategies are more appropriate in a given case (which is necessary to know whether one of these is overgeneralized) has been limited due to the fact that past work has rarely involved explanations that are unambiguous in form. Instead, in most studies, the explanations are open to both atheist and agnostic interpretations, thus making both strategies potentially appropriate (see Table 2). For example, many studies provided explanations that did not mention the absence of other causes (e.g., asking participants to compare "Cause A" vs. "Causes B and C"; e.g., Lombrozo, 2007; Vrantsidis & Lombrozo, 2022) or that were inconsistent in doing so (e.g., asking participants to compare "Cause A only" vs. "Causes B and C," so that only the simple explanation was clearly in atheist form; Johnson et al., 2019). Because of communication norms, there is an inherent ambiguity in explanations like "Cause A" and "Causes B and C": these could be interpreted as either agnostic about any unmentioned causes, or as implying the absence of causes that are contextually salient (Dulany & Hilton, 1991; Rooy, 2004). Studies in which participants generate and report their own explanations involve similar ambiguities (Bonawitz & Lombrozo, 2012; Walker et al., 2017), unless participants chose to clearly specify how they were thinking of additional causes (which was not coded for in these experiments). Thus, with respect to the question of whether agnostic strategies are overgeneralized to cases for which atheist explanations are more appropriate, we have only identified one set of relevant studies, where explanations were unambiguously specified in atheist form for both simple and complex explanations (Pacer &

Lombrozo, 2017). Yet, as we will explain, the overall pattern of results in these studies does not unambiguously support either an atheist or agnostic strategy.

The second challenge is that it is difficult to distinguish whether people used atheist or agnostic strategies in previous work, since most work has not examined cases for which the predictions of these two strategies clearly diverge (e.g., where one predicts a complexity preference, while the other predicts a simplicity preference; see Figure 1). Instead, the vast majority of existing work has focused on cases for which the predictions either plausibly converge (i.e., could converge under plausible assumptions), or where the predictions are unclear (see Table 2), such that observing particular explanation preferences is not diagnostic of the strategy in use. For example, some studies have used cases for which all three individual causes are similarly rare (Lombrozo, 2007), or might be assumed to be similarly rare (as with three unknown diseases) (Johnson et al., 2019; Lombrozo, 2007, Study 1; Pacer & Lombrozo, 2017, Study 1; Read & Marcus-Newhall, 1993). In such cases, both agnostic and atheist strategies could plausibly converge in producing the observed simplicity preferences. In other cases, both strategies could converge in producing *complexity* preferences – e.g., when B and C are much more common than A, as in other conditions of Bonawitz and Lombrozo (2012), Lombrozo (2007), and Pacer and Lombrozo (2017). Other studies do not provide clear predictions for atheist and agnostic strategies. For example, Vrantsidis and Lombrozo (2022) did not provide all of the information needed to compute predictions for both atheist and agnostic strategies. In particular, while they provided some relevant information (such as prior probabilities for the two explanations, "A" and "B and C"), other information was missing, including the probabilities and dependencies of the *individual* causes ("A," "B," and "C"), as well as knowledge of how participants interpreted the priors – as applying to the atheist or agnostic versions of the

explanations. (Johnson et al., 2019, Study S4 runs into similar issues.) Another case where these two strategies do not make clear predictions is Shimojo et al.'s (2020) studies. These studies only ever mention (or provide prior probabilities for) *one* explanation per scenario (e.g. only "A," or only "B and C") without ever mentioning other possible explanations or causes. This makes it impossible to know what an atheist strategy would predict – since the nature of the absent causes, not to mention their probabilities, are completely unknown.

Despite the prevalence of these challenges, a few existing studies have avoided them: three papers examine conditions in which the predictions of atheist and agnostic strategies clearly diverge (Bonawitz & Lombrozo, 2012; Lombrozo, 2007; Pacer & Lombrozo, 2017), and one of these uses unambiguously atheist explanations (Pacer & Lombrozo, 2017). Yet the specific patterns of results observed make it difficult to infer whether atheist or agnostic strategies were used in these cases. For example, two of these papers included conditions in which the atheist strategy predicts a complexity preference, while the agnostic strategy predicts no preference (Bonawitz & Lombrozo, 2012, 1:1 condition; Pacer & Lombrozo, 2017, 1:1 condition). Yet the observed results in these conditions were not clearly consistent with either strategy, as adult participants (in Lombrozo, 2007) and children (in Bonawitz & Lombrozo, 2012) tended to show *simplicity* preferences – predicted by neither strategy – while adults in the latter study showed complexity preferences – in line with atheist strategies. The third paper (Pacer & Lombrozo, 2017) again included conditions in which atheist and agnostic predictions diverge, and this time provided unambiguously atheist explanations. In the two relevant conditions, the atheist strategy predicted no preference (in the 1:1 condition) or a simplicity preference (in the 1:2 condition), while the agnostic strategy predicted a complexity preference in both cases. (The reversed direction of these predictions reflects the specific causal structure

used in this work, where A could cause B and C.) Yet the results were again not fully consistent with either strategy: in both of these conditions, a simplicity preference was found – thus only partly in line with an atheist strategy, and partly in line with neither strategy. Moreover, across these studies, the observed simplicity preferences seem to reflect a broader pattern of bias towards the simpler explanation, even when both strategies predicted complexity preferences (e.g., because B and C were common). Thus, these results suggest that people may sometimes *over*simplify, with whatever strategy they are using, but cannot clearly distinguish whether an agnostic or atheist strategy led to their simplicity preference in the first place.

Overview of Current Experiments

The current work thus aimed to differentiate agnostic and atheist strategies for evaluating explanations, and to examine whether agnostic strategies are overgeneralized even in cases where an atheist strategy is most appropriate. To do this, we focused on cases where participants were asked to evaluate unambiguously atheist explanations (that clearly specified absences), and where the predictions of the two strategies clearly diverge. More concretely, and returning to our clock example, asking about atheist explanations would mean asking participants to compare the following two explanations for the clock's problems: misplaced spring-plugs, but not worn gears or a winding malfunction (i.e., A, but not B or C), vs. worn gears and a winding malfunction, but not misplaced spring-plugs (i.e., B and C, but not A). Given this wording of the explanations, the most appropriate (i.e., literally correct) interpretation of the question involves treating the explanations as atheist, allowing us to test whether agnostic strategies are overused despite this disambiguation. Moreover, in order to differentiate atheist and agnostic strategies, we used scenarios for which the three causes are *common* (i.e., each cause has an 80% chance of being

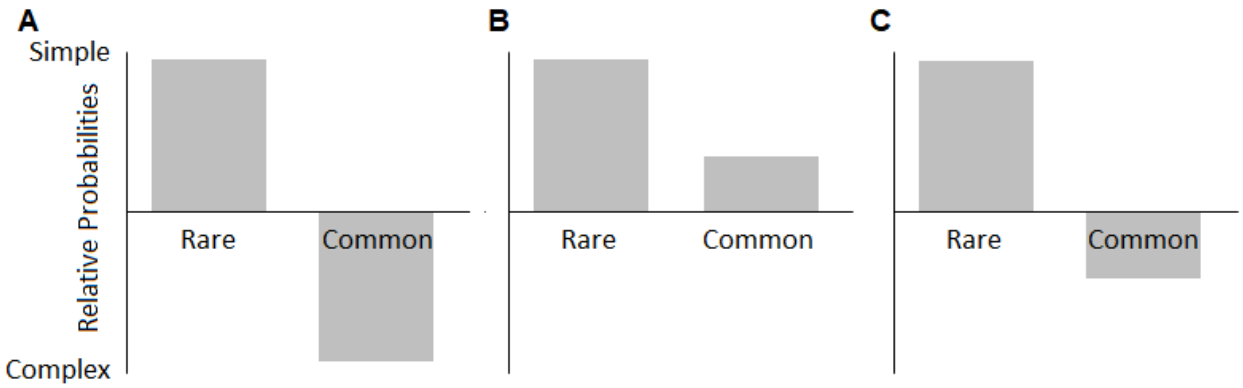
present), such that an agnostic strategy should produce simplicity preferences, whereas an atheist strategy should produce complexity preferences.

While our key predictions concern the case in which causes are common, our studies also include an equivalent condition in which causes are rare (present 20% of the time), to more precisely identify biases towards overgeneralizing agnostic strategies, while accounting for the fact that strategies may differ across participants. Specifically, comparing the 80% and 20% cases allows us to make more precise predictions about the strength of simplicity/complexity preferences that should result from using an atheist strategy, without having to assume that participants perform exact probabilistic calculations. In particular, due to the specific probabilities used in the 20% and 80% cases (which swap the probabilities of a cause being present vs. absent), an atheist strategy should produce a full preference reversal in these two conditions: a symmetrical flip around a mid-point indicating no preference. In other words, we should observe complexity preferences in the common condition that are equal in magnitude to the simplicity preferences observed in the rare condition (see Figure 2A and Table 1). In contrast, an agnostic strategy should produce simplicity preferences in both conditions (though perhaps a weaker simplicity preference in the common condition, based on the ratio of the posteriors shown in Table 1; see Figure 2B). These predictions allow us to identify a tendency to overgeneralize agnostic strategies: if at least some participants inappropriately apply agnostic, rather than atheist, strategies in the current studies, the averaged responses should be shifted away from the atheist predictions, and towards the agnostic predictions. More concretely, this would mean that average responses in the common condition would fail to show a symmetrical flip to complexity preferences, and instead show biases towards simplicity (i.e., such that there are simplicity preferences in both conditions, or weaker complexity preferences in the common

condition as compared to the magnitude of simplicity preferences in the rare condition); see Figure 2C. We refer to this as a *bias* towards simplicity to indicate that the pattern of explanation preferences favors simplicity beyond what is justified based on the available information, which includes the provided probability information (such as the 20% or 80% chance of the causes being present) and the nature of the explanations being asked about (i.e., explanations that are clearly in atheist form).

Figure 2

Qualitative Predictions of Different Strategies in the Current Studies



Note. (A) Predictions if all participants use a purely atheist strategy: explanation preferences should show a full reversal such that the magnitude of simplicity preferences in the rare condition (causes present 20% of the time) is equal to the magnitude of complexity preferences in the common condition (causes present 80% of the time). (B) Predictions if all participants use a purely agnostic strategy: explanation preferences should show simplicity preferences in both conditions, with a potentially weaker preference in the common condition. (C) Predictions if at least some participants use agnostic strategies in the current studies (thus overgeneralizing these strategies to cases for which atheist strategies are more appropriate): average responses in the common condition should fail to show the full preference reversal and instead show a bias

towards simplicity, with results somewhere in between the pattern of results predicted by purely atheist or agnostic strategies (one of a range of possibilities is shown in the figure).

Across three studies and one supplementary study, we thus test whether participants show biases towards simplicity preferences in the common condition (consistent with a tendency to overgeneralize agnostic strategies), or whether participants instead show a symmetrical flip from simplicity to complexity preferences across the rare vs. common conditions (consistent with using the appropriate atheist strategy). To foreshadow our results, Studies 1 and S1 find that participants show biases towards simplicity, in line with overgeneralizing an agnostic strategy. Studies 2 and 3 therefore further test the idea that this is related to overgeneralizing agnostic strategies, and insufficiently considering the absences specified in the explanations, by manipulating how relevant the absences seem, and showing that this attenuates the previously observed simplicity bias. All studies were fully preregistered (see Open Science section for details).

Table 2

Previous Research on Preferences for Simpler vs. more Complex Explanations

Paper	Type of explanation (atheist, agnostic, or ambiguous)	Did predictions of atheist and agnostic strategies diverge?	Key results
Bonawitz & Lombrozo, 2012	Unknown (participant-generated explanations)	In 1:1 condition: Atheist – complexity preference Agnostic – no preference	Across conditions: Simplicity preference in children (beyond what was justified by objective probabilities), except when complex explanation was much more likely.

			Complexity preference for adults (possibly driven by objective probabilities).
Johnson et al., 2019	Atheist for simple, Ambiguous for complex	Unclear	Simplicity preference.
Liefgreen & Lagnado, 2023	Ambiguous	Unclear	Simplicity preference when drew causal graphs. No clear preference otherwise.
Lombrozo, 2007	Ambiguous	In 1:1 condition: Atheist – complexity preference Agnostic – no preference	Across conditions: Simplicity preference (beyond what was justified by objective probabilities), except when complex explanation is much more likely, and/or difficulty of probabilistic computations reduced.
Pacer & Lombrozo, 2017	Atheist	In 1:1 condition: Atheist – no preference Agnostic – complexity preference In 1:2 condition: Atheist – simplicity preference Agnostic – complexity preference	Across conditions: Simplicity preference (sometimes beyond what was justified by objective probabilities), specifically for fewer *unexplained/root* causes.
Read & Marcus-Newhall, 1993	Ambiguous	Unclear	Simplicity preference.
Shimojo et al., 2020	Agnostic	Unclear	Simplicity preference (beyond what was justified by objective probabilities), but only when simplicity was made salient.
Vrantsidis & Lombrozo, 2022	Ambiguous	Unclear	Simplicity preference (beyond what was justified by objective probabilities).

Walker et al., 2017	Unknown (participant-generated explanations)	Unclear	Simplicity preference.
Zemla et al., 2017	Mixed (Study 1), Ambiguous (Study 2)	Unclear	Complexity preference.
Zemla et al., 2023	Ambiguous	Unclear	Variable preference. Weaker simplicity preference/ stronger complexity preferences when mechanism information provided.

Note. The results column notes whether the simpler or complex explanations tended to be preferred, and, for cases where priors and likelihoods were provided or elicited, whether this preference went beyond what was justified by the objective probabilities. This table does not include research that uses different forms of simplicity/complexity, as these do not necessarily correspond to the operationalization in the current paper in terms of the number of causes (or unexplained causes) posited in an explanation. Specifically, the table excludes three papers that used subjective ratings of simplicity/complexity (Lim & Oppenheimer, 2020; Liquin & Lombrozo, 2022; Marsh et al., 2022; see also Zemla et al., 2017) and that most often found complexity preferences, though results vary. It also excludes one paper that used the number and range of free parameters to test "Bayesian Occam's Razor" (Blanchard, Lombrozo, et al., 2018), and observed simplicity preferences.

Study 1

Study 1 was based closely on previous studies using scenarios like the clock example in which participants tended to prefer simpler explanations (e.g., Johnson et al., 2019; Lombrozo, 2007; Vrantsidis & Lombrozo, 2022) – that is, to assign simpler explanations higher posterior

probabilities (i.e., higher probabilities, given the effects to be explained), in addition to assigning them higher prior probabilities (i.e. higher probabilities, prior to observing the effects). Here, we made two key modifications to these studies. First, we provided explanations in clearly atheist form, by explicitly specifying the absence of other causes. Second, we manipulated whether causes were described as rare (present 20% of the time) or common (present 80% of the time). As discussed, this frequency manipulation allows us to more clearly identify the overgeneralization of agnostic strategies, by testing whether or not participants show a symmetrical flip in explanation preferences across frequency conditions. Including the rare condition (for which both atheist and agnostic strategies predict a simplicity preference) also allowed us to ensure that the simplicity preferences found in previous work indeed replicated with our current task and materials.

Note that the structure of the scenarios means that, mathematically, the same predictions hold for both posterior and prior probabilities. This is because the causes were guaranteed to produce their effects (i.e., both the simple and complex explanations have Bayesian likelihoods of 100%). This means that the ratio of the two explanations' priors is equal to the ratio of their posteriors (see Table 1). The current studies therefore assess simplicity preferences using both judgments of posterior and prior probabilities, to ensure that results generalize across both judgments. Moreover, since previous work suggests that simplicity might have additional influences on posteriors that are not mediated through priors or likelihoods (Vrantsidis & Lombrozo, 2022), we also examine any differences in the strength of simplicity preferences across these two judgments.

Methods

Participants

291 adult participants were recruited through Prolific (age: M = 38; gender: 198 women, 77 men, 16 additional or multiple responses). An additional 8 participants were excluded for failing attention checks or answering comprehension checks incorrectly by their second attempt. The preregistered minimum sample size for each study was determined a priori through a combination of power analyses based on pilot data, precedents from previous research in this area, and pragmatic factors (e.g., cost). With our sample size of 291, we had 80% power to detect an effect of frequency condition (common vs. rare) on posteriors of b = .35 (i.e., a difference between conditions of .70 scale points). Power was estimated through simulations. That is, multiple sets of outcomes were simulated at various effect sizes, using the data structure from the main dataset (e.g., number of participants, number of responses per condition), and using variance estimates (i.e., unexplained variance and participant and scenario level variance) from the main analysis for which power was computed.

Design

The key theoretically relevant manipulation in this study was the frequency of the causes, where these causes could be either rare or common (varied randomly across each of a participant's three trials). In addition, the outcome measure was varied between-participant, so that either posteriors or priors were assessed. Additional counterbalanced factors are listed below.

Procedure

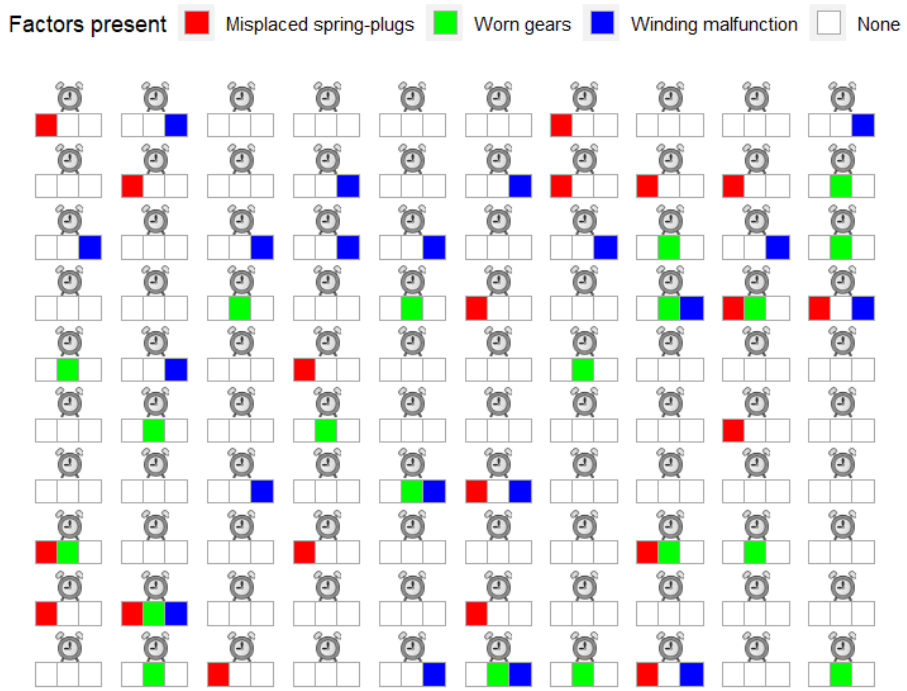
The study involved three trials. In each trial, participants read a scenario and answered several questions about it. The scenarios were always visible for the corresponding questions. Each scenario described two effects that could be observed, with one cause (Cause A) that would produce both effects, and two causes (Causes B and C) that would each produce one effect. In what follows, we focus on two possible explanations for observing the pair of effects, referred to

here as the *simple* explanation (Cause A present, and not Cause B or C) and the *complex* explanation (Causes B and C present, and not Cause A). Scenarios were based closely on those from Johnson et al. (2019). An example is shown below (bolding in original). The other scenarios are included in the Supplementary Materials.

> There is a collection of old clocks at the European History Museum.
> Sometimes the clocks have timekeeping problems such as running slow or skipping hours.
>
> Several factors are known to be able to cause these issues. Specifically:
>
> **Misplaced spring-plugs** always cause **slow running and skipping hours.**
> **Worn gears** always cause **slow running**.
> **A winding malfunction** always causes **skipping hours.**
>
> Nothing else is known to cause a clock to run slow or to skip hours.
>
> All three of these factors, misplaced spring-plugs, worn gears, and winding malfunctions, are quite **[rare/common]** in these clocks.
> In particular, misplaced spring-plugs occur in about **[20%/80%]** of clocks, worn gears occur in about **[20%/80%]** of clocks, and winding malfunctions occur in about **[20%/80%]** of clocks.
>
> The proportion of clocks in which each of these factors is present is shown in the image below.

Figure 3

Example Frequency Image from Study 1



Note that, in the scenarios, causes were guaranteed to produce their effects, such that both explanations had a Bayesian likelihood of 100%. Thus, while the observed effects constrain the possible explanations (since, for example, Cause B on its own would not produce them), the likelihood of observing these effects does not provide a reason for favoring either of the two explanations provided. This means that, mathematically, the explanations' relative posterior probabilities should only depend on their relative prior probabilities. This approach to setting likelihoods has been successfully used in several studies to isolate simplicity's effects through prior probabilities (Johnson et al., 2019; Shimojo et al., 2020).

Across scenarios, the frequency of the causes was varied randomly, such that the causes were either described as rare, and each present in 20% of cases, or as common, and each present in 80% of cases. To emphasize this information, an image that visually displayed these frequencies was shown after each scenario (as in Figure 3), with each cause present in either 20 or 80 out of 100 cases. The frequency with which the displayed causes co-occurred corresponded

to their probability of co-occurring if causes were statistically independent. At the start of the study, participants were familiarized with how to interpret these images, and were given two chances to answer a set of comprehension questions about the images.

After reading a scenario, participants compared the simple and complex explanations, either in terms of posterior or prior probabilities (varied between-participants). For posteriors, participants reported whether they thought the pair of effects was more likely to have been caused by the simple explanation or the complex explanation, by providing a response on a sliding scale from -5 to 5. For example: "One of these clocks was observed to have both slow running and skipping hours. How likely is this to have been caused by: -5 = misplaced spring-plugs (not worn gears or a winding malfunction), 5 = worn gears and a winding malfunction (not misplaced spring-plugs)." Note that the absence of other cause(s) was explicitly specified in each explanation. For priors, participants were asked, e.g., "Imagine that we randomly select one of these clocks. Which of the following types of clocks do you think we are more likely to have selected? One with…" (response scale the same as for posteriors).

Across participants, we counterbalanced whether the causes invoked by the simple (vs. the complex) explanation were described first in the scenarios, images, and questions. Note that the names of the causes were swapped when using the counterbalanced ordering (e.g., so that the simple explanation only involved cause C present), but for clarity, we refer to the causes using the original labelling throughout. Different topics were used for each scenario (broken clocks, UV waves, or soil issues) to increase generalizability.

After completing all three trials, participants saw each scenario again without the frequency information, and reported their beliefs about the frequency of causes in the scenario, based on their pre-existing knowledge: e.g., "Based on your knowledge of the real world, on

average, how rare or common do you think the factors that could cause these problems are (i.e. the factors that could cause slow running or skipping hours)?" Responses ranged from 0, "Extremely rare (occurring in 0% of these clocks)," to 100, "Extremely common (occurring in 100% of these clocks)." This measure was included in case participants relied on these pre-existing beliefs instead of the manipulated frequencies.

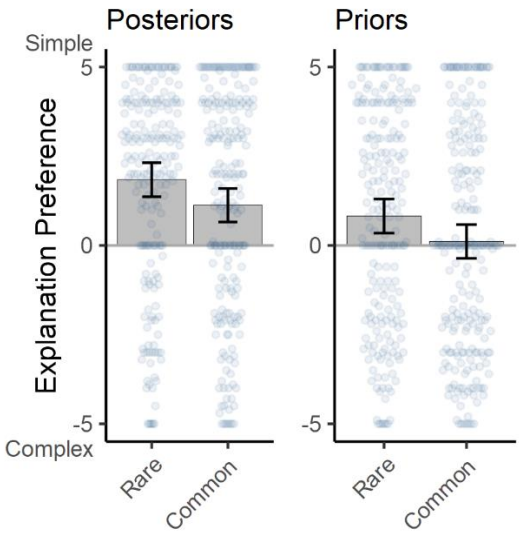
Finally, participants reported demographic information and were debriefed.

Results

Analyses for all studies were performed in R (R Core Team, 2021), using the lmerTest package (Kuznetsova et al., 2017) for multilevel models. For all studies, probability ratings were recoded so that positive values indicated simplicity preferences (i.e., higher probability for the simple than the complex explanation). All Study 1 analyses were fit as multilevel models with random intercepts for participant and scenario.

Figure 4

Simplicity vs. Complexity Preferences in Study 1.



Note. Plot displays explanation preferences (i.e. ratings of explanations' relative probability on a

-5 to 5 scale), where positive values indicate that the simpler explanation was rated as more probable than the complex explanation, and negative values indicate the reverse. 95% CIs shown.

Table 3

Percentage of Participants showing Agnostic and Atheist Response Patterns in Study 1

Response Pattern	Posteriors (% participants)	Priors (% participants)
Agnostic	42.71	24.55
Atheist	13.54	11.82

Note. A participant's response pattern was coded as agnostic if their responses displayed simplicity preferences across all trials, for both the rare and common condition. A participant's response pattern was coded as atheist if all rare condition responses displayed simplicity preferences, while all common condition responses displayed complexity preferences. The remaining participants displayed other response patterns (see Supplementary Materials for full breakdown).

Posteriors

We first examined simplicity/complexity preferences in judgments of posterior probabilities. Judgments of relative posterior probabilities were predicted from frequency condition (1 = common, -1 = rare). As shown in Figure 4, in the rare condition, simpler explanations tended to be seen as more probable (b = 1.84, 95% CI [1.38, 2.31], $p < .001$). The frequency of causes did have a small effect (b = -0.36, 95% CI [-0.60, -0.11], p = .005), such that simplicity preferences were weaker in the common condition. However, participants on average still showed simplicity preferences in this case (b = 1.13, 95% CI [0.67, 1.58], $p < .001$), in line

with an agnostic interpretation of the explanations, even though the atheist form of the provided explanations meant that the complex explanation was in fact more probable.

Consistent with these results, Table 3 shows that more participants responded in line with agnostic strategies – showing simplicity preferences in both conditions – compared to atheist strategies – showing simplicity preferences when causes were rare, and complexity preferences when causes were common (binomial test: $b = .76$, 95% CI [0.62, 0.867], $p < .001$). Note that results involving individual response patterns were not preregistered, and should be considered exploratory for all studies. Because characterizing individual response patterns was not the original goal of this study, the design did not allow for all participants' response patterns to be coded, since, for some participants, all three trials were randomly assigned to the same frequency condition. Results regarding individual response patterns are therefore based only on participants whose responses could be coded ($n = 96$ for posteriors, $n = 100$ for priors; excluded participants: $n = 46$ for posteriors, $n = 39$ for priors).

Priors

Equivalent analyses on prior probabilities yielded similar results. As shown in Figure 4, participants in the rare condition tended to assign simpler explanations higher prior probabilities ($b = 0.82$, 95% CI [0.36, 1.28], $p < .001$). The frequency of causes again had a small effect ($b = -0.36$, 95% CI [-0.61, -0.11], $p = .005$), such that participants in the common condition showed a small, non-significant simplicity preference ($b = 0.11$, 95% CI [-0.34, 0.57], $p = .63$), but again failed to show mean complexity preferences. Again, more participants responded in line with agnostic strategies than with atheist strategies (see Table 3; binomial test: $b = .68$, 95% CI [0.51, 0.81], $p = .03$).

We also compared the effects for priors and posteriors. When predicting responses from

response type (1 = posterior, -1 = prior) and its interaction with cause frequency, there was a main effect of response type (b = 0.51, 95% CI [0.24, 0.78], p < .001), but no interaction with frequency (b = 0.00, 95% CI [-0.17, 0.18], p = .99). This indicates that, overall, simplicity preferences were weaker for judgments of priors, compared to posteriors.

Subjective Frequency of Causes

While the observed pattern of preferences is consistent with participants' tending to use agnostic strategies, an alternative possibility is that participants may have instead used an atheist strategy, but one that was based on their pre-existing beliefs about the causes' frequencies, rather than the provided information. If this were the case, it would predict stronger complexity preferences (or weaker simplicity preferences) to the extent these pre-existing beliefs imply that causes are more common. To test this, two secondary analyses predicted posteriors or priors from reported beliefs about cause frequency (controlling for the provided frequencies). Despite these beliefs being distributed widely across the scale (M = 52.90, SD = 25.81, range = [0,100]), there were no significant effects of these beliefs on judgments of posteriors (b = -0.00, 95% CI [-0.01, 0.01], p = .68) or priors (b = 0.00, 95% CI [-0.01, 0.01], p = .82), suggesting that participants were not merely using an atheist strategy that relied on these pre-existing beliefs.

Discussion

Study 1 replicated prior work in that it found a preference for simpler explanations when causes were rare. Going beyond prior work, Study 1 also found that preferences for simpler explanations were on average attenuated, but not reversed, when causes were common. Importantly, we examined cases for which all explanations were clearly atheist – i.e., they specified the absence of additional causes. Given these atheist explanations, along with the other features of our scenarios (e.g., the 20% vs. 80% frequencies for the causes, and matched

Bayesian likelihoods), the correct atheist response would have been to prefer simpler explanations when causes were rare, but to show an equally strong preference for complex explanations when causes were common. In contrast, if at least some participants incorrectly treated the explanations as agnostic, thus showing simplicity preferences in both conditions, this would bias the average results towards simplicity preferences (or weaker complexity preferences) in the common condition. Thus, the average response pattern in this study is consistent with the hypothesis that many participants overgeneralized an agnostic strategy, despite being asked to compare atheist explanations. Individual participants' response patterns further support this interpretation.

While the findings of Study 1 are consistent with the current hypothesis, we conducted an additional study (Study S1) to ensure that key results were not driven by potential methodological limitations. Study S1, reported in full in the Supplementary Materials, involved three key changes to the methods of Study 1. First, the wording of the posterior questions was clarified to ensure that it was not misinterpreted as asking about causal *responsibility* (which might lead to favoring simpler explanations, if a single cause that produces two effects is seen as more causally powerful or responsible). Specifically, the posterior question was changed to ask directly about the probability of the explanation given the effects, without using causal language. For example, in the clock scenario, it asked: "One of these clocks was observed to have both slow running and skipping hours. How likely is this clock to have each of the following combinations of factors?" Second, rather than evaluating only two explanations ("A & not B or C" and "B & C & not A"), participants evaluated all eight possible combinations of causes being present/absent (no causes present, only Cause B present, etc.). This ensured that the results of Study 1 were not an artifact of focusing on only the simple and complex explanations. It also

highlighted that participants were being asked to evaluate *atheist* explanations, since the question separately asked about all of the possibilities that would be subsumed under an agnostic explanation (e.g., for the simple explanation, it asked separately about "A & not B & not C," "A & B & not C," "A & C & not B," and "A & B & C"). Third, the visual frequency images from Study 1 were removed, to confirm that results were not driven by perceptual biases that might affect how frequencies were inferred from these images.

Using these updated methods, Study S1 replicated the key results of Study 1, finding mean simplicity preferences when causes are rare, and either simplicity preferences (for posteriors) or no significant preference (for priors) when causes are common. This replication can increase confidence in the reliability of Study 1's results, and help ensure that the observed bias towards simplicity was not driven by the specific methodological choices of Study 1.

Study 2

The results of Studies 1 and S1 are consistent with the hypothesis that people tend to apply agnostic strategies in evaluating explanations, and in fact over-generalize these strategies, applying them even when explanations are specified in atheist form. However, the observed pattern of simplicity preferences could also be produced by any number of other factors that might bias people towards simpler explanations: for example, implicitly treating causes as rare, treating causes as negatively dependent such that co-occurrences are especially unlikely, or preferring simpler explanations for non-probabilistic reasons (e.g., because they are easier to process, Vrantsidis & Lombrozo, 2022; Wilkenfeld, 2019). Studies 2 and 3 therefore aimed to provide more direct evidence for our proposed mechanism: the tendency to use agnostic strategies for evaluating explanations, i.e,, strategies that only consider the causes that are specified as present in an explanation, and thus fail to represent or appropriately consider those

causes posited to be absent.

Study 2 examined this idea by testing whether making absences more relevant would shift responses towards an atheist pattern of responding. In particular, we manipulated whether, within a given explanation, the causes that were specified as absent were causally relevant to the effects being explained, in the sense that changing those causes from absent to present (without changing the status of the other causes in the explanation) would have made a difference to whether the effects occurred (Strevens, 2004; see also Vasilyeva et al., 2018; Woodward, 2010). In the scenarios used in Study 1, the causes that were absent in the explanations were causally *irrelevant* in this sense, since the causes already specified as present in the explanations were sufficient to produce the effects, and changing the absent cause(s) to present would not have further altered those effects. For example, within the simple explanation, the presence of A plus the *absence* of B and C would produce the same effects as the presence of A plus the *presence* of B and C. Participants may have thus ignored the absent causes, treating the provided atheist explanations as merely agnostic. In Study 2, scenarios in the "irrelevant" condition were structured to maintain this property, similar to Study 1. In contrast, in the "relevant" condition of Study 2, the scenarios were changed so that the presence or absence of the absent causes *did* make a difference to the explanandum (e.g., so that A, B, and C simultaneously being present would *not* produce the observed effects). We expected that when absences were causally relevant in this way, participants would be more likely to consider absent causes.[3] Therefore, if the

[3] There is another complementary way to understand this manipulation: as varying the *stability* with which the agnostic explanations will produce their effects (i.e., varying whether their ability to produce their effects depends on other moderating factors (Blanchard, Vasilyeva, et al., 2018; Vasilyeva et al., 2018; Woodward, 2010). Concretely, in the "irrelevant" condition, Cause A will stably produce its effects, regardless of the status of Causes B and C, while in the "relevant" condition, the effects of Cause A are unstable, and depend on the status of B and C (equivalently for Causes B and C depending on A). In contrast, the set of causes described in the atheist explanations, which includes both present and absent causes, will stably produce its effects across both conditions. Since people tend not to like unstable explanations (Vasilyeva et al., 2018), this is another interpretation of why the "relevant" condition might shift participants from agnostic to atheist explanations.

previously observed bias towards simplicity was indeed driven by using an agnostic strategy that in some way involved ignoring these absences, this manipulation should push responses closer to the atheist pattern of responding: i.e., flipping from simplicity preferences when causes are rare to complexity preferences when causes are common. This finding would thus provide more direct support for the role of agnostic strategies in driving the results observed in Studies 1 and S1.

Including the "relevant" condition of Study 2 has an additional benefit. In this condition, only the atheist explanations (which specify absences) can adequately account for the observed effects, since the agnostic versions would not produce the same effects if the additional, unspecified, causes turned out to be present. Thus, if we continue to find a bias towards simplicity in the "relevant" condition, this would provide another form of evidence that agnostic strategies tend to be overgeneralized even when atheist strategies are more appropriate in the sense that only the atheist explanation can properly account for the observed effects.

Finally, Study 2 also examined participants' self-generated explanations, which were elicited before giving them specific explanations to compare. These self-generated explanations can offer insight into the strategies people use in their spontaneous explanatory reasoning, as reflected in the pattern of simplicity/complexity preferences found in these explanations, as well as whether the explanations explicitly mention causes being absent.

Methods

Participants

389 adult participants were recruited through Prolific (age: M = 38; gender: 178 women, 201 men, 10 additional or multiple responses). Twelve additional participants were excluded for failing comprehension checks by their second attempt. The final sample size provided 80%

power to detect an effect size of $b = 0.33$ for the interaction of frequency condition and relevance condition on posteriors. This effect size is equivalent to the difference between the rare and common condition changing by 1.32 scale points across the relevant vs. irrelevant conditions. Power was computed through simulation as in Study 1.

Design

The key theoretically-relevant manipulations formed a 2 (absent causes relevant vs. irrelevant; between-participant) x 2 (frequency of causes: common vs. rare; within-participant) design. All participants completed all outcome measures (posteriors, priors, and open-ended explanations). Additional counterbalanced factors are listed below.

Procedure

Participants were assigned either to a condition in which absent causes were *irrelevant* to the observed effects (like Study 1), or *relevant* to the observed effects (unlike Study 1). Four pairs of scenarios were created (about machines, medication, coffee drinking, or massage appointments), with an irrelevant and a relevant version of each. The machines scenario is used here as an example, and the other scenarios are included in the Supplementary Materials.

As in Study 1, all scenarios described some effects that could be observed, and a set of three causes that could produce the effects. In the relevant condition scenarios, there were only two ways to produce the observed effects: Cause A being present and not Cause B or C (the simple explanation), or Causes B and C being present and not Cause A (the complex explanation), as in the following scenario:

> There are three machines in a factory: Machine A, Machine B, and Machine C.
> If a machine is running low on oil, a light will turn on, indicating that the oil needs to be refilled.
>
> When Machine A's light is on, it needs to be refilled with **4 ounces** of oil.
> When Machine B's light is on, it needs to be refilled with **2 ounces** of oil.

> When Machine C's light is on, it needs to be refilled with **2 ounces** of oil.
>
> [The machines are not used frequently, so the oil does not need to be refilled very often./ The machines are used frequently, so the oil needs to be refilled quite often.] Specifically, each machine needs to be refilled on about [**20%/80%**] of days.
> The amount that each machine is used, and thus whether it needs to be refilled on a given day, is unrelated to how much the other machines are used.

The observed effect in this scenario was that "One day, exactly 4 ounces of machine oil had been used from the store-room." This could be produced in only two ways: by refilling Machine A, and not B or C, or by refilling Machines B and C, and not Machine A. Importantly, this is because the absent causes here are causally relevant, in that, if present, they would change the observed effects (e.g., refilling all three machines would use eight ounces of oil, not four).

In contrast, in the irrelevant condition, making these absent causes present would *not* change the observed effects. Using the machine scenario as an example, participants in the irrelevant condition instead learned:

> When Machine A's light is on, it needs to be refilled with both **mineral oil** and **synthetic oil**.
> When Machine B's light is on, it needs to be refilled with **mineral oil**.
> When Machine C's light is on, it needs to be refilled with **synthetic oil**.

In this case, the observed effects were that "One day, both the mineral oil and synthetic oil had been used from the store-room." This could be produced by the simple explanation (Machine A was refilled, and not B or C), the complex explanation (Machines B and C were refilled, and not A), or any version of these where the absent causes were made present (e.g., all three machines were refilled).

Participants read two such scenarios, and answered several questions based on each with the corresponding scenario visible for reference. One scenario was presented in the rare condition (where causes were described as occurring infrequently and as being present in only 20% of cases), and the other was presented in the common condition (where causes were

described as occurring frequently and as present in 80% of cases). To ensure that participants paid attention to this frequency information, participants had two chances to correctly answer a comprehension question that required reporting the frequency of causes (20%/80%). As in Study S1, frequency images were not included. Instead, a sentence indicating the independence of the three causes was added to each scenario (e.g., see the last sentence in the machine scenario above).[4]

For each scenario, participants then provided an open-ended explanation for the observed effects. For example, in the relevant condition machine scenario, they were asked:

> In a sentence or two, please write down what you think BEST EXPLAINS why exactly 4 ounces of machine oil had been used from the store-room, in terms of which of the machines had been refilled that day. Please write only what you think is the **SINGLE BEST EXPLANATION**, even if you think there are multiple possible explanations. Please be **as specific as possible** in your response.

(Capitalization and bolding in original.) Following this, participants were given the opportunity to check their explanation by learning whether particular causes were present or absent. This was intended as an additional measure of whether participants were considering absent causes (i.e., we expected participants who were agnostic about absent causes not to check for their presence or absence). However, this measure did not work as intended; further details and discussion are

[4] It is possible that some participants may have missed this sentence, or not understood that the causes were meant to be independent of each other. However, misunderstandings of this type are unlikely to have affected the results of this study, for several reasons. First, results in the 'irrelevant' condition replicate those in Study 1 (where the images should have made the independence quite salient), and those in Study S1 (which assessed whether participants understood that causes were independent, and only included participants who understood this correctly). There is also no obvious reason that participants' independence assumptions should be altered in the 'relevant' conditions. Second, the content of the scenarios was intentionally designed so that the independence of the causes was plausible, given reasonable real-world expectations, making this misunderstanding perhaps less likely. Third, if this misunderstanding occurred, the most plausible way for causes to be related in these scenarios is for their status as present or absent to be positively correlated, such that they co-occur above chance (e.g., if there are busy periods in the factory, when all three machines are used frequently). Using agnostic explanations, this assumption should still lead to a simplicity preference, though a smaller one (since the combination of B and C being present is now more likely than when causes are independent), as was found by Pacer and Lombrozo, (2017). Using atheist explanations, this assumption would not necessarily affect preferences at all (since both explanations involve two causes occurring in the same state – present/absent – and one in a different state). Thus, the qualitative predictions of using atheist vs. agnostic explanations are not altered.

reported in the Supplementary Materials.

Participants then reported relative posterior and prior probabilities for the simple vs. complex explanations. For posteriors, the question was similar to Study 1, but used non-causal language, as in Study S1: e.g., "One day, exactly 4 ounces of machine oil had been used from the store-room. Which is more likely?" 5 = "Machine A was refilled that day, but not Machine B or Machine C, 5 = "Machine B and Machine C were refilled that day, but not Machine A." Priors were elicited as in Study 1. An additional exploratory measure was included after the posterior and before the prior ratings, where participants reported whether each cause was more likely to be present vs. absent given the observed effects. Results of this measure are reported in the Supplementary Materials.

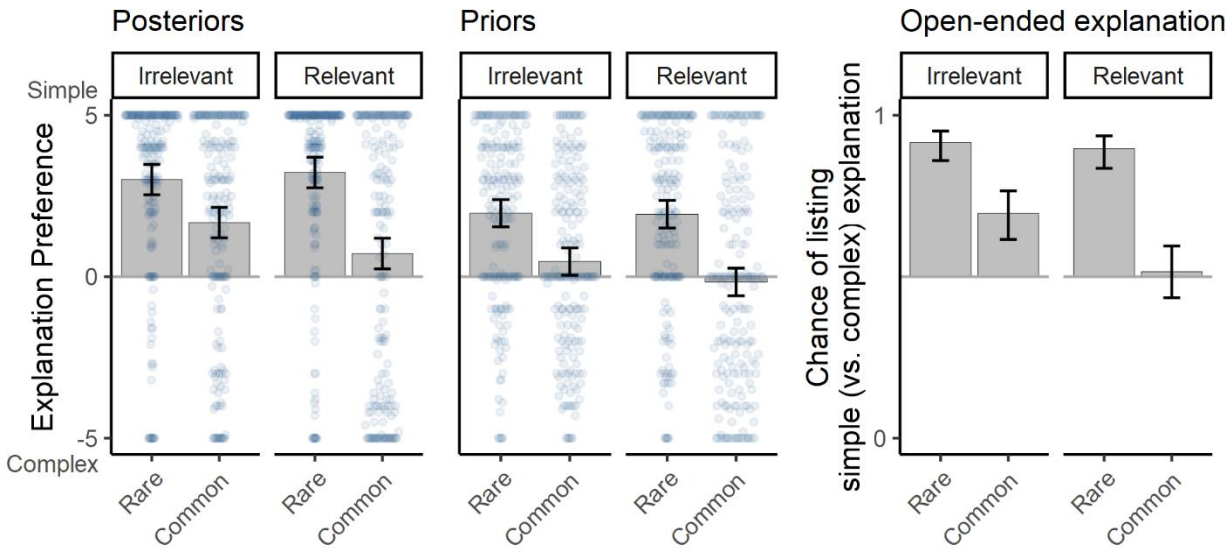
Across participants, we counterbalanced whether the causes involved in the simple (vs. the complex) explanation were described first, and whether the rare vs. common condition scenario was presented first. A random two out of four scenario topics were used for each participant.

Results

All analyses in Studies 2 and 3 were fit as multilevel models with random intercepts for each scenario topic. Mediation was tested using the lavaan package (v. 0.6-9; Rosseel, 2012) in R.

Figure 5

Simplicity vs. Complexity Preferences across the Relevant and Irrelevant Conditions in Study 2



Note. For posteriors and priors, plots display explanation preferences (i.e., ratings of explanations' relative probability on a -5 to 5 scale), where positive values indicate that the simpler explanation was rated as more probable than the complex one, and negative values indicate the reverse. For open-ended explanations, values above 0.5 indicate a greater chance of listing the simple explanation compared to the complex explanation; values below 0.5 indicate the reverse. 95% CIs shown.

Table 4

Percentage of Participants showing Agnostic and Atheist Response Patterns in Study 2, by Relevance Condition

Response Pattern	Posteriors (% participants)		Priors (% participants)	
	Irrelevant Condition	Relevant Condition	Irrelevant Condition	Relevant Condition
Agnostic	63.27	51.81	42.35	34.20
Atheist	16.33	34.20	18.88	31.09

Note. A participant's response pattern was coded as agnostic if their responses displayed simplicity preferences in both the rare and common condition. A participant's response pattern

was coded as atheist if they displayed a simplicity preference in the rare condition, and a complexity preference in the common condition. The remaining participants displayed other response patterns (see Supplementary Materials for full breakdown). The percentages shown were computed within each relevance condition.

Posteriors

We first examined how posterior probability judgments varied across conditions, by predicting posterior judgments from the relevance condition (1 = relevant, -1 = irrelevant), interacted with the frequency of causes. There was a significant interaction (b = -0.29, 95% CI [-0.52, -0.07], p = .01; see Figure 5). Breaking this down, results in the irrelevant condition replicated Study 1: participants tended to show simplicity preferences in both the rare (b = 3.23, 95% CI [2.77, 3.69], p < .001) and common (b = 0.72, 95% CI [0.26, 1.18], p = .002) conditions, and these simplicity preferences were weaker when causes were common (frequency effect: b = -1.25, 95% CI [-1.58, -0.93], p < .001). This pattern also held in the relevant condition (rare: b = 3.01, 95% CI [2.55, 3.46], p < .001; common: b = 1.67, 95% CI [1.22, 2.13], p < .001; frequency effect: b = -0.67, 95% CI [-0.99, -0.34], p < .001). However, as predicted, making absent causes relevant (vs. irrelevant) further weakened simplicity preferences when causes were common (b = -0.48, 95% CI [-0.80, -0.15], p = .004), though not enough to produce a mean complexity preference. Thus, increasing the relevance of absent causes pushed responses more towards the predictions of an atheist strategy, though not enough to produce a full flip to complexity preferences, suggesting that agnostic strategies continued to be overused. Individual response patterns supported these results (see Table 4). In both conditions, more participants responded in line with agnostic rather than atheist strategies (binomial test: irrelevant: b = .79, 95% CI [0.72,

0.86], $p < .001$; relevant: $b = .60$, 95% CI [0.52, 0.68], $p = .01$). Moreover, logistic regressions predicting each specific response strategy from relevance condition (1 = relevant, -1 = irrelevant) showed that, in the relevant (vs. irrelevant) condition, there was a decrease in agnostic responses ($b = -0.23$, 95% CI [-0.44, -0.03], $p = .02$), and an increase in atheist responses ($b = 0.48$, 95% CI [0.25, 0.74], $p < .001$), suggesting that the relevance manipulation shifted participants' responses more towards atheist rather than agnostic strategies.

Priors

Using the same analysis for priors as for posteriors, the pattern of results was qualitatively similar (see Figure 5). However, smaller effect sizes meant that not all predicted effects reached significance. In particular, when looking across both frequency conditions, the relevance manipulation did not generate a significant main effect ($b = -0.17$, 95% CI [-0.37, 0.04], $p = .12$) nor interaction with frequency ($b = -0.15$, 95% CI [-0.36, 0.06], $p = .15$). However, an exploratory analysis within just the common condition (where the relevance effect was predicted) found that relevance did have a significant effect ($b = -0.32$, 95% CI [-0.61, -0.02], $p = .03$): as with posteriors, increasing the relevance of absent causes further reduced simplicity preferences. Also mirroring posteriors, simplicity preferences were found when causes were rare (in the relevant condition: $b = 1.93$, 95% CI [1.52, 2.35], $p < .001$; in the irrelevant condition: $b = 1.96$, 95% CI [1.55, 2.37], $p < .001$), and these were weaker when causes were common (frequency effect in the relevant condition: $b = -1.05$, 95% CI [-1.34, -0.76], $p < .001$; in the irrelevant condition: $b = -0.76$, 95% CI [-1.04, -0.46], $p < .001$). However, in neither condition was there a full flip to mean complexity preferences when causes were common. Instead, in the irrelevant-common condition there was a mean *simplicity* preference ($b = 0.46$, 95% CI [0.06, 0.87], $p = .03$), while in the relevant-common condition there was a small, non-

significant complexity preference (b = -0.16, 95% CI [-0.58, 0.25], p = .44), with an exploratory test confirming that this was of a much smaller magnitude than the simplicity preferences observed in the relevant-rare condition (difference in magnitude (rare minus common) = 1.77, 95% CI [1.18, 2.36]). These results again suggest that participants had a tendency to overgeneralize agnostic strategies, which was somewhat, but not completely, reduced when absences were relevant.

Individual participants' response patterns also supported this idea. There were more agnostic than atheist responses in both conditions, though this difference was only significant for the irrelevant condition (binomial test: irrelevant: b = .69, 95% CI [0.72, 0.86], p < .001; relevant: b = .52, 95% CI [0.52, 0.68], p = .52). Moreover, atheist strategies became more common (b = 0.33, 95% CI [0.25, 0.74], p = .006), and agnostic strategies became marginally less common (b = -0.17, 95% CI [-0.44, -0.03], p = .10), in the relevant compared to the irrelevant condition (tested as for posteriors).

Because some effects of the relevance manipulation were significant for posteriors but not for priors, an exploratory analysis tested if these results significantly differed from each other. Responses were predicted from the interaction of relevance, frequency, and response type. Neither the main effect of relevance (b = -0.01, 95% CI [-0.16, 0.14], p = .90), nor its interaction with frequency (b = -0.07, 95% CI [-0.22, 0.08], p = .36), significantly differed for priors vs. posteriors. Though these differences were not significant, the numerically smaller size of the relevance effects observed for priors may reflect the fact that evaluating prior probabilities requires considering only causes, and not effects, while evaluating posteriors requires considering both. Thus, manipulating the relevance of absent causes to these effects likely had less impact on prior probability judgments. In addition, this analysis found a significant main

effect of response type (b = 0.55, 95% CI [0.40, 0.71], p < .001), which showed that there were overall stronger simplicity preferences for posterior judgments compared to priors, replicating Study 1.

Open-ended explanations

Examining participants' open-ended explanations allowed us to test whether similar patterns of agnostic vs. atheist strategies occurred in participants' spontaneous explanatory reasoning. Explanations were coded by two coders in terms of whether each of the three causes was mentioned as present, mentioned as absent, or was unmentioned, or alternatively, whether the explanation should be excluded for not fulfilling the criteria of the question (e.g., responses listed multiple explanations, or effects were not explained in terms of the three potential causes). For subsequent analyses, explanations that could not produce the observed effects were also excluded, as they may reflect inattentive responses. The two coders had 81% agreement. Analyses in Study 2 are based on the first coder's ratings, but all results replicated with the second coder's ratings (i.e., all significant results remained significant and in the same direction, while non-significant results remained non-significant).

Using these open-ended explanations, exploratory analyses examined how the relevance condition affected simplicity/complexity preferences: roughly, whether the generated explanations invoked the presence of a single cause vs. multiple causes. These results replicated the overall pattern of results found for posterior and prior ratings (see Figure 5, and Supplementary Materials for details). This suggests that participants' spontaneous explanatory reasoning tended to default to agnostic strategies in the irrelevant condition, but was pushed somewhat towards atheist strategies in the relevant condition.

The format of the open-ended explanations could also provide another indicator of the

type of explanation being considered (agnostic or atheist). Specifically, if participants were considering atheist explanations that involve absent causes, then they may be more likely to specify the absence of causes in their open-ended explanations (e.g., saying "Machine A *and not Machines B or C*", or perhaps "Machine A *only*", versus merely saying "Machine A" and mentioning nothing about the status of other causes). Mentioning absent causes (1 = yes, 0 = no) was predicted from relevance condition in a multilevel logistic regression. Supporting the idea that the relevance manipulation increased the use of agnostic explanations, participants were more likely to mention absent causes in the relevant, vs. the irrelevant, condition (23% vs. 4% of coded explanations, respectively; $b = 0.97$, 95% CI [0.67, 1.30], $p < .001$). Note that, overall, only a small proportion of responses mentioned absences, perhaps because communication norms for concision often lead people to omit this information, even if they are mentally considering absent causes (Dulany & Hilton, 1991; Rooy, 2004).

We also tested whether mentioning absent causes mediated the effect of relevance condition on simplicity preferences, measured through relative posteriors. Deviating from the preregistration (to simplify this analysis), this was tested only in the common condition, where relevance effects were predicted and found. The indirect effect was not significant ($b = -0.12$, 95% CI [-0.60, 0.37], $p = .64$), which was due to the non-significant path from mentioning absent causes to posteriors ($b = -0.14$, 95% CI [-0.77, 0.47], $p = 0.64$), though results were in the predicted direction. These non-significant results may reflect low power due to the small proportion of explanations that mentioned absent causes.

Discussion

Study 2 found that making absences causally relevant (in the sense that specifying their presence vs. absence in an explanation would change the effects produced) shifted responses

more towards a pattern consistent with atheist, rather than agnostic, strategies: further attenuating (but still not fully reversing) preferences for simpler explanations when causes were common. This was found both when participants evaluated explanations provided in atheist form, and when participants generated their own explanations. The results of the relevance manipulation provide more direct support for the role of agnostic strategies in the observed bias towards simplicity, since making absences causally relevant decreased this simplicity bias, likely by increasing the chance that participants would consider the absent causes involved in the atheist explanations. Furthermore, the lack of a symmetrical flip to complexity preferences is again consistent with the hypothesis that simplicity preferences in these scenarios often arise from treating explanations as agnostic, and that this tends to be overgeneralized even to cases where atheist explanations are more appropriate – here, because explanations were provided in atheist form, or because atheist explanations best account for the observations (as in the relevant condition).

Study 3

The results of Studies 1 and 2 support the idea that simplicity preferences can be driven by people's tendency to use agnostic strategies for evaluating explanations, even when atheist strategies are more appropriate. And while Study 2 suggested that it is possible to shift participants towards using atheist strategies, results were not fully in line with the predictions of an atheist strategy, instead continuing to be biased towards simplicity even when absences were causally relevant. One interpretation of this result is that there is a strong default tendency for people to ignore absent causes and treat explanations as agnostic, and that the relevance manipulation was not strong enough to fully overcome this default. Alternatively or additionally, it could be that other factors unrelated to considering absent causes contribute to the bias towards

simplicity observed in Studies 1 and 2.

Study 3 thus aimed to provide a stronger test of the hypothesis that the bias towards simplicity preferences observed in Studies 1 and 2 was driven by a failure to consider absent causes – i.e., by overgeneralizing an agnostic strategy to atheist cases. Therefore, in addition to using causal structures for which absences were causally relevant (as in Study 2), Study 3 introduced a version of each scenario in which the absent causes were replaced with alternative causes that produce their own effects. For example, in the "absence" version of the machine scenario, the causes were described as different machines being on (where the absence of being on – being off – produced no effect). In contrast, in the "alternative cause" version of this scenario, this absence was replaced with an alternative cause, by describing the machine as being in one of two modes (high-power and low-power modes) which each produce different effects (using some different, non-zero, amount of power). This alternative presentation format should encourage participants to explicitly consider these absent/alternative causes in their reasoning, consistent with previous research using similar manipulations (e.g., where people reasoned better about false positives on a mammogram test when the false positives were described as due to the *presence* of benign cysts, rather than merely the *absence* of cancer; Krynski & Tenenbaum, 2007, though see also Hayes et al., 2016, 2018; McNair & Feeney, 2014, 2015). Thus, if this manipulation succeeds in generating a symmetrical flip to complexity preferences when causes are common, it would further suggest that a failure to consider absent causes, and not other factors, is what drove the previously observed bias towards simplicity.

Methods

Participants

Participants were 302 adults recruited through Prolific (age: $M = 35$; gender: 158 women,

136 men, 8 additional or multiple responses). One additional participant was excluded for failing comprehension checks by their second attempt. Due to a delay in registering study completions, the pre-registered sample size of 300 was exceeded by two participants. The final sample size provided 80% power to detect an effect size of $b = .38$ for the interaction of frequency condition and absence condition on posteriors. This effect size is equivalent to the difference between the rare and common condition changing by 1.52 scale points across the absence vs. alternative cause conditions. Power was computed through simulation as in Study 1.

Design

The key theoretically-relevant manipulations formed a 2 (absence vs. alternative cause condition; between-participant) x 2 (frequency of causes: common vs. rare; within-participant) design. As in Study 2, all participants completed all outcome measures (posteriors, priors, and open-ended explanations). Counterbalanced factors were also the same as in Study 2.

Procedure

Participants were assigned to one of two conditions: the "absence" condition, which was based on the "relevant" condition from Study 2, or the "alternative cause" condition, in which absent causes were replaced with an alternative cause that would produce different effects. Two scenarios from Study 2 (machines and medicine) were modified to create matched scenarios in each condition. The machine scenarios are presented here, and medicine scenarios are included in the Supplementary Materials. In the alternative cause condition, the machine scenario was as follows (bolding in original):

> There are three machines in a factory: Machine A, Machine B, and Machine C.
> The factory is open 24-7. During that time, each machine is sometimes in high-power mode and sometimes in low-power mode.
>
> If Machine A is in **high-power** mode, it uses **100 watts** of power.
> If Machine A is in **low-power** mode, it uses **10 watts** of power.

> If Machine B is in **high-power** mode, it uses **50 watts** of power.
> If Machine B is in **low-power** mode, it uses **5 watts** of power.
>
> If Machine C is in **high-power** mode, it uses **50 watts** of power.
> If Machine C is in **low-power** mode it uses **5 watts** of power.
>
> The machines are [rarely/usually] in high-power mode. Specifically, each machine is in high-power mode about **[20%/80%]** of the time.
> Whether a machine is in high- or low-power mode at a given time is unrelated to whether the other machines are in high- or low-power mode at that time.

Here, the observed effect was that "at this moment, a total of 110 watts of power is being used by the machines," and this could be explained by two possible explanations: "Machine A is in high-power mode, and Machines B and C are in low-power mode" or "Machines B and C are in high-power mode, and Machine A is in low-power mode." Note that we continue to refer to these as the simple and complex explanations, respectively, despite the fact that they both involve the presence of three causes. We use this nomenclature to facilitate comparisons with the simple and complex explanations used in our other conditions and studies.

In contrast, the 'absence' version of the machine scenario was as follows:

> There are three machines in a factory: Machine A, Machine B, and Machine C.
> The factory is open 24-7. During that time, each machine is sometimes on and sometimes not.
>
> If Machine A is on, it uses **100 watts** of power.
>
> If Machine B is on, it uses **50 watts** of power.
>
> If Machine C is on, it uses **50 watts** of power.
>
> The machines are [rarely/usually] on. Specifically, each machine is on about **[20%/80%]** of the time.
> Whether a machine is on at a given time is unrelated to whether the other machines are on at that time.

Analogous to the alternative cause condition, the observed effect was that *100 watts* of power were being used by the machines, and this could be explained by two possible explanations:

"Machine A is on, and Machines B and C are not" or "Machines B and C are on, and Machine A is not." Thus the key difference between conditions is that, in the alternative cause condition, the scenario explicitly discusses two modes for the machines, and each mode produces different effects. In contrast, in the absence condition, the scenario primarily discusses one mode for the machines (being on), and the alternative (not being on) is not given a separate label, nor described as producing any effect.

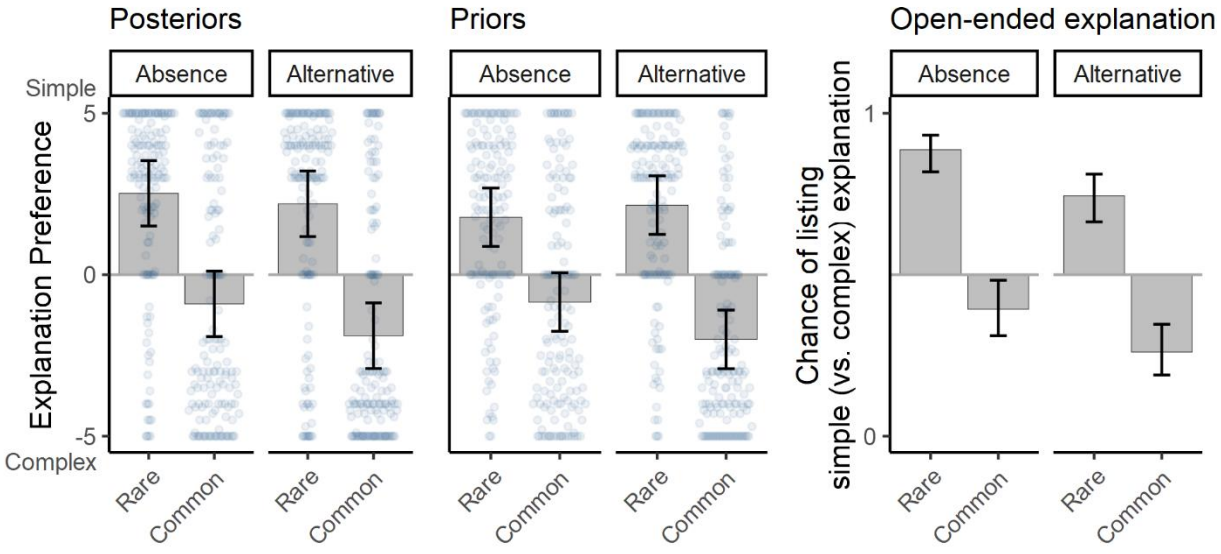
The methods for Study 3 were otherwise the same as those for Study 2, aside from removing the checking measure and the presence/absence judgements used in Study 2.

Results

As in the previous studies, priors and posteriors were recoded so that positive values always indicate preferences for the simpler explanation (in the absence condition), or for the explanation that is analogous to the simple explanation (in the alternative cause condition). For ease of exposition, we refer to these as reflecting simplicity or complexity preferences in both conditions, despite the fact that both explanations in the alternative cause condition posit the presence of the same number of causes, and so are technically equally simple.

Figure 6

Simplicity vs. Complexity Preferences across the Absence and Alternative Cause Conditions in Study 3



Note. For posteriors and priors, plots display explanation preferences (i.e., ratings of explanations' relative probability on a -5 to 5 scale), where positive values indicate that the simpler explanation was rated as more probable than the complex explanation, and negative values indicate the reverse. For open-ended explanations, values above 0.5 indicate a greater chance of listing the simple explanation compared to the complex explanation; values below 0.5 indicate the reverse. 95% CIs shown.

Table 5

Percentage of Participants showing Agnostic and Atheist Response Patterns in Study 3, by Absence vs. Alternative Cause Condition

Response Pattern	Posteriors (% participants)		Priors (% participants)	
	Absence Condition	Alternative Cause Condition	Absence Condition	Alternative Cause Condition
Agnostic	28.67	19.08	24.67	12.50
Atheist	48.67	55.26	42.00	57.89

Note. A participant's response pattern was coded as agnostic if their responses displayed simplicity preferences in both the rare and common condition. A participant's response pattern

was coded as atheist if they displayed a simplicity preference in the rare condition, and a complexity preference in the common condition. The remaining participants displayed other response patterns (see Supplementary Materials for full breakdown). The percentages shown were computed within each relevance condition.

Posteriors

We first examined whether simplicity preferences in posteriors depended on condition. Specifically, responses were predicted by the alternative cause vs. absence condition (-1 = absence, 1 = alternative cause), interacted with whether causes were rare or common (see Figure 6). The interaction was not significant (b = -0.16, 95% CI [-0.43, 0.10], p = .22), though it was in the predicted direction, and there was a significant main effect of alternative cause vs. absence condition (b = -0.33, 95% CI [-0.59, -0.07], p = .01). This indicates that there were weaker simplicity preferences in the alternative cause condition, with Figure 6 showing that this difference was numerically larger when causes were common (in line with predictions).

Despite this non-significant interaction, exploratory analyses broke the results down by alternative cause vs. absence condition. In the absence condition, the results were expected to mirror the relevant condition in Study 2. And, indeed, the pattern of results was largely similar. Specifically, the simplicity preferences found when causes were rare (b = 2.52, 95% CI [1.83, 3.20], p = .003) were reduced significantly (b = -1.71, 95% CI [-2.09, -1.34], p < .001) such that there was a non-significant complexity preference when causes were common (b = -0.90, 95% CI [-2.09, -1.34], p = .07). (In contrast, Study 2 showed a *simplicity* preference here.) However, like Study 2, the absence condition did not produce a symmetrical flip to complexity preferences: an exploratory test showed that the slight complexity preference observed when causes were

common was significantly weaker than the simplicity preferences found when causes were rare (difference in magnitude (rare minus common) = 1.62, 95% CI [0.67, 2.56]). Thus, responses were still biased towards simplicity in this condition.

In the alternative cause condition, the simplicity preferences found when causes were rare (b = 2.19, 95% CI [1.51, 2.88], p = .005) were also significantly reduced (b = -2.04, 95% CI [-2.41, -1.67], p < .001) when causes were common, and, for the first time, this led to a significant *complexity* preference when causes were common (b = -1.89, 95% CI [-2.57, -1.21], p = .008). Moreover, as shown in Figure 6, preferences appeared to show a full, symmetrical flip, and an exploratory test showed that the magnitude of the complexity preference when causes were common was not significantly different from the magnitude of the simplicity preference when causes were rare (difference in magnitude (rare minus common) = 0.31, 95% CI [-0.64, 1.25]). These results imply that replacing absences with alternative causes led participants' mean responses to fall in line with the correct atheist strategy, and no longer show a systematic bias towards simplicity.

Individual response patterns mirrored these results (see Table 5). In both conditions, more participants showed atheist compared to agnostic response patterns (binomial tests: absence: b = .37, 95% CI [0.28, 0.47], p = .007; alternative cause: b = .26, 95% CI [0.18, 0.35], p < .001). The greater prevalence of atheist strategies is consistent with the mean (though sometimes non-significant) complexity preferences found in both conditions when causes were common. However, importantly, responses in the alternative cause (compared to the absence) condition showed a trending decrease in agnostic strategies (b = -0.27, 95% CI [-0.54, 0.00], p = .052), as well as a non-significant increase in atheist strategies (b = 0.13, 95% CI [-0.09, 0.36], p = .25), tested with logistic regressions as in Study 2. Though not quite significant, this result is

consistent with the idea that the experimental manipulation in Study 3 shifted responses from agnostic to atheist strategies.[5]

Priors

Results for priors largely replicated those for posteriors. The equivalent analysis found a significant interaction between the absence vs. alternative cause condition and cause frequency (b = -0.38, 95% CI [-0.61, -0.15], p = .001), in the predicted direction (see Figure 6). Breaking this down, in the absence condition, the pattern of results was the same as in the relevant condition of Study 2. Specifically, the simplicity preferences found when causes were rare (b = 1.77, 95% CI [1.17, 2.39] p = .007) were reduced (b = -1.31, 95% CI [-1.65, -0.98], p < .001) when causes were common, such that there was a non-significant complexity preference (b = -0.85, 95% CI [-1.46, -0.23], p = .06), though an exploratory test showed that this was of a smaller magnitude than the simplicity preference when causes were rare (difference in magnitude (rare minus common) = 0.93, 95% CI [0.09, 1.77]).

In the alternative cause condition, the simplicity preferences when causes were rare (b = 2.15, 95% CI [1.55, 2.76], p = .004) were also significantly reduced (b = -2.08, 95% CI [-2.41, -1.75], p < .001), such that there was a significant complexity preference when causes were common (b = -2.00, 95% CI [-2.61, -1.39], p = .005). And again, as shown in Figure 6, in the

[5] To see whether this manipulation completely removed the bias towards using agnostic strategies when assessing posteriors, we compared the prevalence of agnostic patterns to that of the next most common incorrect response pattern (i.e., showing complexity preferences in both the rare and common condition; see Supplementary Materials). In both conditions, agnostic response patterns were still more common than the next most common pattern (absence condition: 28.67% vs. 8.67%, binomial test: b = .77, 95% CI [0.64, 0.87], p < .001; alternative cause condition: 19.08% vs. 12.50%; binomial test: b = .26, 95% CI [0.45, 0.74], p < .001). This suggests that there was still some tendency to overgeneralize agnostic strategies in the alternative cause condition, beyond other types of errors. However, this tendency appears to have been somewhat reduced in the alternative cause condition. Specifically, a logistic regression predicting response pattern (1 = agnostic, 0 = next most common strategy) from absence vs. alternative cause condition showed a trending decrease in the use of agnostic strategies relative to the next most common strategy (b = -0.38, 95% CI [-0.82, 0.03], p = .07).

alternative cause condition, these complexity preferences were similar in magnitude to the simplicity preferences when causes were rare (difference in magnitude (rare minus common) = 0.15, 95% CI [-0.69, 0.99]).

Individual response patterns also mirrored results for posteriors (see Table 5). In both conditions, more participants showed the appropriate atheist response pattern, compared to agnostic response patterns (binomial test: absence: $b = .37$, 95% CI [0.28, 0.47], $p = .01$; alternative cause: $b = .18$, 95% CI [0.11, 0.26], $p < .001$), with even more atheist ($b = 0.32$, 95% CI [0.09, 0.55], $p = .005$) and fewer agnostic ($b = -0.41$, 95% CI [-0.73, -0.12], $p = .007$) responses observed in the alternative cause compared to the absence condition (tested using logistic regressions as in Study 2).[6]

An exploratory analysis that included the three-way interaction with response type (posterior or prior) found no significant interactions involving response type (ps > 0.22). The main effect of response type found in the previous studies (with weaker simplicity preferences for priors relative to posteriors) was also not significant in this study ($b = 0.11$, 95% CI [-0.07, 0.28], $p = .24$), although the numerical difference across response type was in the same direction.

Open-ended explanations

Secondary analyses examined participants' open-ended explanations. Explanations were coded as in Study 2. The two coders had 74% agreement. Reported results were based on the first coder's ratings, but all results replicated using the second coder's ratings.

[6] While some participants still displayed agnostic response patterns in each condition, in the absence condition this occurred somewhat more frequently than the next most common strategy, which was showing complexity preferences in both conditions (24.67% vs. 12.67% of participants; binomial test: $b = .66$, 95% CI [0.52, 0.78], $p = .02$). In contrast, these frequencies were quite similar in the alternative cause condition (12.50% vs 9.21% of participants; binomial test: $b = .58$, 95% CI [0.39, 0.75], $p = .49$), although the difference between conditions was not significant ($b = -0.18$, 95% CI [-0.62, 0.26], $p = .42$). This tentatively suggests that in the alternative cause condition, the systematic bias to overuse agnostic strategies was largely if not completely reduced, as it occurred at similar rates to other incorrect strategies. See Supplementary Materials for details.

As shown in Figure 6, the pattern of simple vs. complex explanations generated replicated the results for posterior ratings, though, in this case, the complexity preference found when causes were common in the absence condition reached statistical significance (vs. being non-significant for posteriors). Also mirroring posteriors, complexity preferences when causes were common were stronger in the alternative cause condition compared to the absence condition, and only in the alternative cause condition were they similar in magnitude to the simplicity preferences found when causes were rare. Thus, participants' spontaneous explanatory reasoning largely mirrored their evaluations of the provided explanations. (See Supplementary Materials for detailed results.)

An exploratory analysis also examined whether participants' explanations were more likely to mention absent/alternative causes (e.g., mentioning that a machine was off, or in low-power mode) in the alternative cause condition, compared to the absence condition. The response coding and analysis were the same as in Study 2. And indeed, there was a much higher prevalence of mentioning absent/alternative causes in the alternative-cause condition, compared to the absence condition (99.61% vs. 24.28% of included responses, $b = 3.34$, 95% CI [2.57, 4.78], $p < .001$).

An exploratory mediation tested whether mentioning absent/alternative causes mediated the effect of alternative-cause vs. absence condition on simplicity preferences, measured in terms of posteriors. This analysis was restricted to cases in which causes were common. The indirect effect was marginally significant ($b = -2.84$, 95% CI [-5.78, 0.10], $p = 0.06$), and involved a marginally significant path from mentioning absent/alternative causes to posteriors, ($b = -0.79$, 95% CI [-1.61, 0.03], $p = 0.06$), with mentioning absent/alternative causes linked to stronger complexity preferences.

Discussion

Study 3 found that replacing absences with alternative causes eliminated the bias towards simplicity, and produced strong mean complexity preferences when causes were common. These complexity preferences were about equal in magnitude to the simplicity preferences when causes were rare. This is in line with the symmetrical flip in preferences that would follow from using the correct atheist strategy in this task. Note that these results in the alternative cause condition should not technically be considered "complexity" preferences, since both explanations posit the presence of three causes, and thus are of equal simplicity/complexity. Nevertheless, varying the way in which absent causes were presented provides further evidence that how participants thought about absent causes was an important driver of our results across Studies 1 to 3. In particular, it suggests the systematic bias towards simplicity when causes were common was due to a tendency to overgeneralize agnostic strategies that do not consider absent causes, even in cases where these absences are specified in the explanations (as in Studies 1-3), or required to best explain the effects (as in the "relevant" condition of Study 2, and both conditions of Study 3).

Study 3 also helps rule out some alternative explanations for the simplicity bias observed in our previous studies and in the absence condition. Specifically, the symmetrical flip in explanation preferences observed in the alternative cause condition helps rule out any source of systematic bias towards one explanation that would still apply in this condition. For example, this speaks against a systematic tendency to implicitly assume that all causes are rare, or to assume that causes are anti-correlated and thus unlikely to co-occur, since these factors are unlikely to be affected by replacing absences with alternative causes. Of course, we cannot infer that *no* other factors contributed to the observed simplicity biases. For example, people might

still prefer simpler explanations because they are easier to process (Vrantsidis & Lombrozo, 2022; Wilkenfeld, 2019), something which would *not* have applied in the alternative cause condition (due to the equal number of causes involved in the two explanations). Nevertheless, the combined results of Studies 2 and 3 show that failing to consider absent causes – and thus overgeneralizing agnostic strategies for evaluating explanations – is an important driver of biases towards simpler explanations in scenarios like those we test in Studies 1-2 and the absence condition of Study 3.

General Discussion

People often prefer simpler explanations, i.e., those that posit the presence of fewer causes (e.g., Lombrozo, 2007; Read & Marcus-Newhall, 1993; Vrantsidis & Lombrozo, 2022; see also Pacer & Lombrozo, 2017). The current work aimed to distinguish between two possible mechanisms that could drive this preference: using "agnostic" vs. "atheist" strategies for evaluating explanations. As defined here, agnostic strategies are those for which people act as if they are reasoning over agnostic explanations (e.g., "Cause A" or "Causes B and C"), which specify the presence of certain causes, but remain neutral about the presence/absence of other causes. In contrast, atheist strategies involve acting as if one is reasoning over atheist explanations (e.g., "Cause A, and not B or C" or "Causes B and C, and not A"), which also specify the absence of other causes. A series of three studies and one supplementary study supported the hypothesis that simplicity preferences tend to arise from agnostic strategies for evaluating explanations, and that people tend to over-generalize this agnostic strategy to cases in which an atheist strategy is more appropriate, by in effect ignoring the absences posited by an explanation.

The explanation preferences found in Study 1 and Study S1 were consistent with a tendency to overgeneralize agnostic strategies even when asked about explicitly atheist explanations, a case for which an atheist strategy would be more appropriate. Specifically, while participants on average preferred simpler explanations when causes were rare (i.e., assigning them higher prior and posterior probabilities), they did not show a symmetrical flip to preferring complex explanations when causes were common, as predicted by an atheist strategy. Instead, mean responses still showed a bias towards simpler explanations when causes were common (e.g., showing simplicity preferences, or weak complexity preferences), in line with the influence of an agnostic strategy. Examining individual participants' responses further showed the predominance of agnostic over atheist response patterns.

Studies 2 and 3 provided further evidence that this bias towards simplicity was indeed driven by over-generalizing agnostic strategies, and thus failing to consider the absent causes involved in an explanation. In particular, manipulations that increased participants' consideration of absent causes (making absences causally relevant to the effects, or describing absences as alternative causes) shifted responses towards greater consistency with atheist rather than agnostic strategies. This was observed in mean results, through an increase in complexity preferences when causes were common, which, in most cases examined, was mirrored in individuals' response patterns. Studies 2 and 3 also replicated these results with participants' self-generated explanations, indicating that the observed strategies are also reflected in participants' spontaneous explanatory reasoning. In addition, the "relevant" and "absence" conditions of Studies 2 and 3, respectively, showed that agnostic strategies were overgeneralized in a different sense – i.e., applied even in scenarios for which absences were required to produce the effects, providing a different sense in which atheist explanations were most appropriate. Thus, overall,

these results suggest that people tend to default to evaluating explanations using agnostic strategies, even when atheist strategies are most appropriate, though this default can be partially overcome (e.g., by making absences more causally relevant). This work sheds light on some of the mechanisms driving people's (sometimes over-generalized) preferences for simpler explanations.

Why might Agnostic Strategies be Overgeneralized?

One question that arises from these findings is *why* people might default to using agnostic strategies for evaluating explanations: what function (if any) might this serve, and what more specific cognitive mechanisms might give rise to it? There are at least three possibilities. First, agnostic strategies might reflect a cognitively efficient way to simplify one's representations: reducing the number of entities that need to be represented by at least temporarily excluding absences from one's working causal model (see Fernbach et al., 2010, for related results). In real-world situations, this strategy might be particularly important, since there could be an unlimited number of potential causes for some observed effect. For example, suppose you want to explain why Alice got a glass of water. One possible explanation is that Alice was thirsty. While representing this single-cause explanation might be fairly easy, it might become cognitively intractable to represent this in an atheist form that explicitly stipulates the absence of every other possible cause (e.g., "Alice was thirsty, and Alice was not trying to water her plants, or put out a fire, or examine water's refractive properties…"). While of course people could represent a more concise version of this atheist explanation (e.g., "Alice was thirsty, and there was no additional reason for her getting water"), they might instead default to the even more concise agnostic explanation that leaves out all absences ("Alice was thirsty"). Because comparing either agnostic or atheist explanations often leads to the same result (see Figure 1),

people might default to this simplest agnostic form unless an atheist form is obviously required in a given case. If people in fact simplify their representations in this way, this suggests that ignoring absent causes might be one strategy people use to address the variable selection problem in causal representation: i.e., the problem of selecting which of all possible candidate causes to represent (Henne et al., 2017; Hesslow, 1988; Kinney & Lombrozo, 2022). This could complement other strategies for variable selection, such as focusing on factors that violate norms or expectations (Gerstenberg & Stephan, 2021; Henne et al., 2017).

It is also possible that the default to use agnostic strategies arises not from people's representations, but from their inference strategies. For example, whether or not absences are represented, people might use cognitively efficient heuristics that only require them to consider the number of causes that are *present* (such that absences are in effect ignored). Indeed, some participants explicitly mentioned using such an inference strategy. For example, one participant in Study 2 said that the complex explanation was less likely because it would involve "two separate events as opposed to one single event" – suggesting the use of a "two events are less likely than one" heuristic, where only present causes count as events. Another participant in Study 3 reported thinking of the saying, "All things being equal, the simplest explanation is usually the correct one" – an explicit use of Ockham's razor, presumably applied by comparing the number of causes that are present, and ignoring the number that are absent. Despite sometimes producing errors, such heuristics might be used because they are often correct – e.g., when causes are similarly rare – regardless of whether agnostic or atheist explanations are involved (see Figure 1). Moreover, these types of heuristics may have other benefits, such as being easier or more reliable to compute than alternative inference strategies which involve combining the probabilities of multiple causes. This increased ease or reliability might be

especially important when faced with atheist explanations, where there could be many absent causes to consider (e.g., Alice not trying to water her plants, not putting out a fire, etc.), or where the probabilities of these absences might be difficult to estimate (e.g., the probability of "all other causes" being absent, when those causes are unspecified) – thus making it especially difficult to compute the joint probability of these absent causes. If people indeed use these types of simplicity-based heuristics, this would be consistent with the broader idea that "explanatory virtues" like simplicity or breadth can serve as heuristics for approximating explanations' probabilities (Glymour, 2015; Johnson et al., 2019; Lipton, 2004; Lombrozo, 2016; Mackonis, 2013; Thagard, 1978, 1989; Vrantsidis & Lombrozo, 2022; Wojtowicz & DeDeo, 2020). While the current work cannot distinguish between heuristic-based and representation-based reasons for defaulting to agnostic strategies, future work can endeavor to distinguish these two possibilities – perhaps by using subtler memory or reaction time tasks to examine whether absences are in fact represented by default (similar to Henne et al., 2017), or by using inference tasks that do not require participants to compare multiple explanations, such that comparative heuristics like Ockham's razor might be less applicable (as in Shimojo et al., 2020).

Yet another possible reason why people might default to agnostic strategies stems from the difficulties of encoding atheist explanations – though this reason does not fit quite as well with the overall pattern of results in the present work. In our studies, participants might not have fully encoded the absent causes when reading the provided explanations – i.e., ignoring the "not B or C" or "not A" – so that they mistakenly thought they were being asked about agnostic explanations, rather than atheist. This idea is in line with linguistic research on the difficulty of processing negations. For example, phrases like "not tall" are sometimes harder to process than "short," at least when not facilitated by one's context or pre-existing schemas (Mayo et al., 2004;

Orenes et al., 2016; Wang et al., 2021). In the current work, this encoding difficulty could have affected participants' responses to the provided explanations (i.e., affecting posterior and prior ratings). However, the fact that similar results were found for self-generated explanations suggests that difficulties encoding negations were not the primary driver of our results. Nevertheless, in other contexts, these encoding factors could play a role in overgeneralizing agnostic explanations. Moreover, it is worth considering whether some shared underlying process might make it more difficult to both encode linguistic negations, and to represent or reason using absent causes. For example, if participants' self-generated explanations are represented verbally, and difficulties *encoding* negations also extend to *generating* negations, this could bias people towards using agnostic, rather than atheist, explanations. Future work could examine this possibility by comparing verbal tasks like those in the current work to non-verbal equivalents (e.g., visual- and motor-based tasks), or by interfering with verbal processing.

Additional Factors that might drive Simplicity Preferences for Posteriors, compared to Priors

The results from comparing posteriors and priors in the current studies can provide additional clues to the mechanisms driving simplicity preferences, but they also raise additional questions. In particular, Studies 1, S1, and 2 (but not 3) found that simplicity preferences were stronger when rating explanations' posterior probabilities, compared to their prior probabilities. While this difference was not predicted by our theoretical framework (since priors and posteriors should be mathematically equivalent in the current scenarios), it nevertheless mirrors a similar result in previous work (Vrantsidis & Lombrozo, 2022). One explanation for this result comes from this previous work, which revealed two distinct uses of simplicity: directly using simplicity as a cue to posteriors, as well as using simplicity as a cue to priors or likelihoods, which in turn

indirectly influence posteriors. As a result, simplicity appears to influence priors only once, but posteriors twice: through both a direct and an indirect effect. This could lead to stronger simplicity preferences for posteriors than priors, as observed in the current work.

Our interpretation of the current findings also suggests other possible reasons for the stronger simplicity preferences on posteriors. For example, if people are relying on inference heuristics such as "all things being equal, the simplest explanation is usually the correct one," it is possible that this heuristic is used more often when the options being compared are viewed as *explanations* – i.e., because they relate causes to effects, as in posterior judgments, rather than just involving causes, as in prior judgments. It is also possible that posterior judgments produce a stronger tendency to exclude absent causes from one's representations, since, when people need to relate causes and effects (rather than to just consider causes), they might focus their representations specifically on causes that are seen as more causally relevant to, or responsible for, those effects (cf. Fernbach et al., 2010). Either of these latter two possibilities would suggest that the simplicity biases observed in these studies not only reflect domain general heuristics (e.g., for comparing probabilities of conjunctions of events), but may also reflect factors that are more specific to causal or explanatory reasoning. Future work is needed to test these ideas more directly.

Do People Select Between Agnostic vs. Atheist Strategies, or Combine Them?

While the current work suggests that agnostic strategies are often overused even when atheist strategies are more appropriate, there are still open questions about the precise nature of this overuse. In particular, the current work has largely assumed that individuals select between atheist and agnostic strategies, and that people overuse agnostic strategies by selecting these more often than they should (e.g., sometimes selecting them even when asked about explanations

in atheist form). However, it is also possible that, rather than selecting between these strategies, participants use a blended strategy that combines both (with the weight put on each strategy potentially varying in different conditions). If these blended strategies exist, then the observed overuse of agnostic strategies might instead reflect an overly high weight placed on the agnostic aspect of a blended strategy. One way that such blended strategies could occur is if people are uncertain about whether the atheist or agnostic causal structure holds, and thus mentally represent both possibilities, assigning each a prior probability of applying in the current case. Then explanation evaluations could reflect a combination of the posteriors inferred from each structure, weighted by each structure's prior (see, e.g., Körding et al., 2007 for a similar model of perceptual inference). If this type of blended strategy was in fact used in our studies, it would mean that the overuse of agnostic strategies was not serving as an energy-saving heuristic, but instead as a computationally effortful way to account for uncertainty about causal structures. Future work can test whether such blended strategies might have indeed driven results in the current studies. Testing this will likely require participants to complete multiple trials in a given condition, and more careful modelling of individual participants' results, in order to distinguish consistent use of a blended strategy from either noisy implementation of a fixed (atheist or agnostic) strategy, or switching between these two strategies on a trial-to-trial basis.

Implications for Predicting Biases towards Simpler Explanations

The current results have implications for predicting when people might be biased towards simpler explanations (i.e., favor simpler explanations in ways that are unwarranted by the available information, including both the provided probability information, and the explanations that are being asked about, or that are most relevant in the current context). By focusing on the tendency to over-generalize agnostic strategies as one source of such biases, the current work

suggests that simplicity biases may occur in cases where the following two conditions hold: where using agnostic strategies predicts greater simplicity preferences than using atheist strategies, and where using atheist explanations is more appropriate.

Indeed, there are a range of conditions under which agnostic strategies produce greater simplicity preferences than atheist strategies (see Figure 1). The most straightforward cases are those examined in the current studies: when all causes are equally common (> 50% chance of being present). While the current scenarios were admittedly artificial, there are many real-world cases that have analogous structures. For example, in regions where poverty is prevalent, the factors that cause poverty (e.g., poor quality education, unaffordable housing) may each have a high chance of being present for any individual person. Moreover, as shown in Figure 1B, a much wider range of cases can produce similarly diverging predictions. Thus, for example, even if the factors that cause poverty are not exactly equal in probability (nor fully independent, etc.), these divergences could still occur.

There are also many real-world cases in which atheist explanations are more appropriate than agnostic explanations. The current studies largely focused on one way in which atheist explanations can be more appropriate: because absences are explicitly mentioned as part of the explanation. And this likely occurs in various real-world situations. For example, atheist explanations may often be raised in the context of a disagreement (e.g., between a prosecution and defense lawyer in court, between competing scientific perspectives, or between conspiracy theories and mainstream views). In such cases, the explainer might want to both specify the absence of causes invoked by competing explanations (e.g., specifying that the defendant did *not* commit the crime), along with putting forward their own proposed cause (e.g., that someone else did commit the crime). In other cases, such as those used in Studies 2 and 3, the absences

specified in an atheist explanation might be required to best account for the observed effects, making the atheist explanation more appropriate in a different sense. As a real-world example of this situation, consider cooking: the result depends not only on the ingredients included, but also on the absence of other ingredients (e.g., making a vanilla cake requires not including chocolate in it). In this case, even if the absence of other ingredients is not explicitly specified in a recipe, this interpretation is most appropriate as it is the only one that will produce the correct results. Finally, explanations that specify the status of all potential causes (rather than remaining agnostic about some) might be especially relevant when intervening to prevent or produce some effect. For example, suppose someone's symptoms could be caused by Disease A and/or Disease B. Knowing that the person has Disease A is not necessarily enough for proper treatment; ideally, one would want to know if the person also has Disease B, in order to treat both if they are both present. And empirical work supports this broad line of reasoning: people's everyday explanations do sometimes mention absences (Zemla et al., 2017), and people do sometimes view absences and omissions as "causes" (Henne et al., 2017, 2019; Wolff et al., 2010), further suggesting that reasoning over atheist explanations can be important in our daily lives.

Importantly, however, to make real-world predictions about when people will be biased towards simpler explanations, it will also be important to examine how the current results generalize to a wider range of situations, as well as to other populations. Here we highlight a few of the factors that might vary in more naturalistic situations, and how these might affect the simplicity biases observed here.

First, it will be important to consider what happens when varying two of the additional probabilistic assumptions used in the current work: the assumption that causes are independent, and that causes are guaranteed to produce their effects. If there is reason to believe that these

assumptions do not hold, this could potentially change what the mathematically-justified preference is, and so what counts as *over*simplification (i.e., a probabilistically unjustified bias towards simplicity, given the available information). For example, sometimes causes are positively dependent, so that co-occurrences are especially common (e.g., the factors that cause poverty may tend to mutually reinforce each other). If there is reason to believe these dependencies hold, then, mathematically, the simpler explanation should be even *less* probable, so that the same simplicity preference would count as even more of a bias. People's actual simplicity/complexity preferences may also shift when these assumptions are changed. For example, people sometimes shift towards complexity preferences when causes are not guaranteed to produce effects (Johnson et al., 2019; though see Vrantsidis & Lombrozo, 2022). This may occur especially when multiple causes can additively increase the chance of an effect occurring, such as when multiple risk factors increase the chance of having a disease. This was plausibly the case in many studies that observed complexity preferences while using more naturalistic explanations (e.g., Lim & Oppenheimer, 2020; Liquin & Lombrozo, 2022; Zemla et al., 2017). Thus, predicting people's real-world explanatory preferences, and understanding when these preferences reflect biases such as oversimplification, will require considering the role of these additional probabilistic factors.

To predict real-world errors in explanatory reasoning, it will also be important to consider people's goals for their explanations. The current work focused on cases where the goal is to select the most probable cause for some effects. Other explanatory goals might shift simplicity/complexity preferences in various ways (Aronowitz & Lombrozo, 2020; Zemla et al., 2023). For example, in certain communicative contexts, people may want explanations that are detailed and informative, which likely contributed to the complexity preferences observed in

some previous work (e.g., Liquin & Lombrozo, 2022; Zemla et al., 2023). Alternatively, in other communicative contexts, people may want explanations that are concise and easy to represent, which could lead them to favor simpler explanations (Kinney & Lombrozo, 2022). However, while these goals may shift the overall magnitude of simplicity/complexity preferences, we suspect that these shifts might operate relatively independently from the mechanisms proposed here, which depend on the tendency to avoid representing or reasoning over absent causes.

Similarly, the type of simplicity involved could also affect explanatory preferences. The current work operationalized simplicity as having a smaller number of causes posited as present (following Lombrozo, 2007) or, more precisely, a smaller number of unexplained causes posited as present (Pacer & Lombrozo, 2017). In contrast, other studies have sometimes relied on subjective ratings of simplicity, which may also reflect factors like the number of details in an explanation, or the number of mechanisms by which a cause can influence the effects. These studies often find that people show complexity preferences (though results vary), suggesting that the direction of these preferences might depend on the type of simplicity involved (Lim & Oppenheimer, 2020; Liquin & Lombrozo, 2022; Marsh et al., 2022; Zemla et al., 2017). Moreover, the type of simplicity involved could potentially affect whether and how the proposed mechanisms – based on failures to consider absent causes – apply. In some cases, these mechanisms might apply in similar ways. For example, the current findings might generalize to other forms of simplicity that are defined based on the number of entities posited to exist: e.g., how many causal relationships, or types of causal factors, exist (Baker, 2022; Lu et al., 2008; Walker et al., 2017). This is because representing or reasoning over fewer entities might be cognitively easier, whatever those entities may be, and, in general, non-present or non-existent entities might tend to be less relevant, and thus tend to be ignored. On the other hand, it is not

clear how the current mechanisms would apply to other forms of simplicity (e.g., simplicity understood in terms of having fewer "free parameters" or less precise hypotheses; Blanchard, Lombrozo, et al., 2018). Understanding how the current mechanisms extend to or interact with other metrics for simplicity is an important question for future research.

Future research will also be necessary to understand if and how the current results generalize to other populations, including in other countries beyond the United States. For example, a large body of research suggests that people from East Asian cultures may in some ways prefer more complex explanations, in that they are more likely to attribute causal responsibility to both individuals' dispositions *and* their situations, rather than just dispositions (Choi et al., 1999). Future work can examine if such tendencies translate into a more general tendency to prefer complex explanations with multiple causal factors, and whether this also extends to an increased chance of considering *absent* causes, rather than only those that are present.

In addition to better understanding when errors of oversimplification are likely to occur, there are also open questions about what interventions are most effective in reducing this error. The current work highlights some potential interventions. We found that participants could overcome a default agnostic strategy when absences were more causally relevant, and especially when absences were framed as alternative causes. This suggests that preferences for simplicity could be reduced by interventions that highlight absences as indeed being "causes," "events," or "difference-makers." And there are likely a variety of ways to do this. For example, other work suggests that people are more likely to view absences as "causes" when the absence differs from norms about what typically happens, or what should have happened – e.g., Billy not watering the plant is seen as causing the plant's death especially if Billy *typically* waters the plant, or *is*

supposed to water the plant (Clarke et al., 2015; Gerstenberg & Stephan, 2021; Henne et al., 2017; Wolff et al., 2010; compare to Henne et al., 2019). This suggests that highlighting how an absence differs from expectations or norms could be helpful for reducing the simplicity biases found here (though interestingly, in the current work, participants in the common condition seemed to ignore absences despite these absences being described as rare, which should make them unexpected or atypical). Future work can also explore other interventions to reduce the observed simplicity bias, and compare their effectiveness and practical applicability. These might include asking people to explicitly consider the probability of any absences (e.g., asking separately about the chance that the defendant did *not* commit the crime, and about the chance that someone else did), or teaching generalizable strategies (e.g., teaching people the limitations of heuristics like Ockham's razor).

Conclusion

In summary, the current work sheds light on one of the mechanisms underlying people's preferences for simpler explanations: a tendency to reason as if they are using agnostic explanations (i.e., which only include causes posited as present, and remain uncommitted to the presence/absence of other causes), even when atheist explanations (that specify the absence of other causes) are more appropriate. This work paves the way for more nuanced examinations of the cognitive representations and inference processes that give rise to simplicity and complexity preferences. It also provides additional insights into why people are sometimes biased towards oversimplified explanations, and how people's explanatory reasoning might be improved to more fully appreciate the complexity of the world.

Acknowledgements

We would like to thank members of the Concepts and Cognition Lab and Niv Lab at Princeton University for their helpful feedback on this research.

Declarations

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of Interest/Competing Interests

The authors declare that they have no conflicts of interest.

Ethics Approval

Ethics approval for this research was obtained from the Princeton University ethics committee. The procedures used in this study adhere to the tenets of the Declaration of Helsinki.

Consent to Participate

For all studies, informed consent was obtained from participants before starting the study.

Consent for Publication

Not applicable.

Availability of Data and Materials

All data is available in the Open Science Framework repository (see Open Science section).

Code Availability

All code is available in the Open Science Framework repository (see Open Science section).

Open Science

All studies were fully preregistered, including the hypotheses, design, sample size, analysis plan, and exclusion criteria. Any exploratory analyses and deviations from the preregistrations have been noted. Pre-registrations, materials, data, and analysis scripts for all studies, as well as Supplementary Materials, are available at: https://osf.io/tvz7r/?view_only=cdd0825cb2134684849cfab762ffa3e0.

References

Aronowitz, S., & Lombrozo, T. (2020). Experiential Explanation. *Topics in Cognitive Science*, *12*(4), 1321–1336. https://doi.org/10.1111/tops.12445

Baker, A. (2022). Simplicity. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2022 Edition). <https://plato.stanford.edu/archives/sum2022/entries/simplicity/>

Blanchard, T., Lombrozo, T., & Nichols, S. (2018). Bayesian Occam's razor is a razor of the people. *Cognitive Scienc PsyArXiv e*, *42*(4), 1345–1359.

Blanchard, T., Vasilyeva, N., & Lombrozo, T. (2018). Stability, breadth and guidance. *Philosophical Studies*, *175*(9), 2263–2283. https://doi.org/10.1007/s11098-017-0958-6

Bonawitz, E. B., & Lombrozo, T. (2012). Occam's rattle: Children's use of simplicity and probability to constrain inference. *Developmental Psychology*, *48*(4), 1156.

Choi, I., Nisbett, R. E., & Norenzayan, A. (1999). Causal attribution across cultures: Variation and universality. *Psychological Bulletin*, *125*(1), 47–63. https://doi.org/10.1037/0033-2909.125.1.47

Clarke, R., Shepherd, J., Stigall, J., Waller, R. R., & Zarpentine, C. (2015). Causation, norms, and omissions: A study of causal judgments. *Philosophical Psychology*, *28*(2), 279–293.

Dulany, D. E., & Hilton, D. J. (1991). Conversational implicature, conscious representation, and the conjunction fallacy. *Social Cognition*, *9*(1), 85–110.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2010). Neglect of alternative causes in predictive but not diagnostic reasoning. *Psychological Science*, *21*(3), 329–336.

Gerstenberg, T., & Stephan, S. (2021). A counterfactual simulation model of causation by omission. *Cognition*, *216*, 104842.

Glymour, C. (2015). Probability and the explanatory virtues. *British Journal for the Philosophy of Science*, *66*(3), 591–604.

Hayes, B. K., Hawkins, G. E., & Newell, B. R. (2016). Consider the alternative: The effects of causal knowledge on representing and using alternative hypotheses in judgments under uncertainty. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(5), 723–739. https://doi.org/10.1037/xlm0000205

Hayes, B. K., Ngo, J., Hawkins, G. E., & Newell, B. R. (2018). Causal explanation improves judgment under uncertainty, but rarely in a Bayesian way. *Memory & Cognition*, *46*(1), 112–131. https://doi.org/10.3758/s13421-017-0750-z

Henne, P., Niemi, L., Pinillos, Á., De Brigard, F., & Knobe, J. (2019). A counterfactual explanation for the action effect in causal judgment. *Cognition*, *190*, 157–164. https://doi.org/10.1016/j.cognition.2019.05.006

Henne, P., Pinillos, Á., & De Brigard, F. (2017). Cause by omission and norm: Not watering plants. *Australasian Journal of Philosophy*, *95*(2), 270–283.

Hesslow, G. (1988). The problem of causal selection. In D. J. Hilton (Ed.), *Contemporary science and natural explanation: Commonsense conceptions of causality* (pp. 11–32). Harvester Press.

Johnson, S., Valenti, J. J., & Keil, F. C. (2019). Simplicity and complexity preferences in causal explanation: An opponent heuristic account. *Cognitive Psychology*, *113*, 101222.

Kinney, D., & Lombrozo, T. (2022). Evaluations of Causal Claims Reflect a Trade-Off Between Informativeness and Compression. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44).

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal Inference in Multisensory Perception. *PLOS ONE*, *2*(9), e943. https://doi.org/10.1371/journal.pone.0000943

Krynski, T. R., & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General*, *136*(3), 430–450. https://doi.org/10.1037/0096-3445.136.3.430

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(1), 1–26. https://doi.org/10.18637/jss.v082.i13

Liefgreen, A., & Lagnado, D. A. (2023). Drawing conclusions: Representing and evaluating competing explanations. *Cognition*, *234*, 105382.

Lim, J. B., & Oppenheimer, D. M. (2020). Explanatory preferences for complexity matching. *PloS One*, *15*(4), e0230929.

Lipton, P. (2004). *Inference to the best explanation* (2nd ed.). Oxford University Press.

Liquin, E. G., & Lombrozo, T. (2022). Motivated to learn: An account of explanatory satisfaction. *Cognitive Psychology*, *132*, 101453. https://doi.org/10.1016/j.cogpsych.2021.101453

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*(3), 232–257.

Lombrozo, T. (2016). Explanatory Preferences Shape Learning and Inference. *Trends in Cognitive Sciences*, *20*(10), 748–759. https://doi.org/10.1016/j.tics.2016.08.001

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*(4), 955.

Mackonis, A. (2013). Inference to the best explanation, coherence and other explanatory virtues. *Synthese*, *190*(6), 975–995.

Marsh, J. K., Coachys, C., & Kleinberg, S. (2022). The Compelling Complexity of Conspiracy Theories. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44).

Mayo, R., Schul, Y., & Burnstein, E. (2004). "I am not guilty" vs "I am innocent": Successful negation may depend on the schema used for its encoding. *Journal of Experimental Social Psychology*, *40*(4), 433–449.

McNair, S., & Feeney, A. (2014). When does Information about Causal Structure Improve Statistical Reasoning? *Quarterly Journal of Experimental Psychology*, *67*(4), 625–645. https://doi.org/10.1080/17470218.2013.821709

McNair, S., & Feeney, A. (2015). Whose statistical reasoning is facilitated by a causal structure intervention? *Psychonomic Bulletin & Review*, *22*(1), 258–264. https://doi.org/10.3758/s13423-014-0645-y

Orenes, I., Moxey, L., Scheepers, C., & Santamaría, C. (2016). Negation in context: Evidence from the visual world paradigm. *Quarterly Journal of Experimental Psychology*, *69*(6), 1082–1092.

Pacer, M., & Lombrozo, T. (2017). Ockham's razor cuts to the root: Simplicity in causal explanation. *Journal of Experimental Psychology: General*, *146*(12), 1761.

Pacer, M., Williams, J., Chen, X., Lombrozo, T., & Griffiths, T. (2013). Evaluating computational models of explanation using human judgments. *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*. https://doi.org/10.48550/arXiv.1309.6855

R Core Team. (2021). *R: A language and environment for statistical computing.* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Read, S. J., & Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, *65*(3), 429.

Rooy, R. (2004). Utility of mention-some questions. *Research on Language and Computation*, *2*(3), 401–416.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software*, *48*(2), 1–36.

Shimojo, A., Miwa, K., & Terai, H. (2020). How Does Explanatory Virtue Determine Probability Estimation?—Empirical Discussion on Effect of Instruction. *Frontiers in Psychology*, 3444.

Sober, E. (2006). Parsimony. *The Philosophy of Science: An Encyclopedia*, *2*, 531–538.

Strevens, M. (2004). The causal and unification approaches to explanation unified—Causally. *Noûs*, *38*(1), 154–176.

Thagard, P. (1978). The best explanation: Criteria for theory choice. *The Journal of Philosophy*, *75*(2), 76–92.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, *12*(3), 435–467.

Vasilyeva, N., Blanchard, T., & Lombrozo, T. (2018). Stable causal relationships are better causal relationships. *Cognitive Science*, *42*(4), 1265–1296.

Vrantsidis, T. H., & Lombrozo, T. (2022). Simplicity as a Cue to Probability: Multiple Roles for Simplicity in Evaluating Explanations. *Cognitive Science*, *46*(7), e13169.

Walker, C. M., Bonawitz, E., & Lombrozo, T. (2017). Effects of explaining on children's preference for simpler hypotheses. *Psychonomic Bulletin & Review*, *24*(5), 1538–1547.

Wang, S., Sun, C., Tian, Y., & Breheny, R. (2021). Verifying negative sentences. *Journal of Psycholinguistic Research*, *50*(6), 1511–1534.

Wilkenfeld, D. A. (2019). Understanding as compression. *Philosophical Studies*, *176*(10), 2807–2831.

Wojtowicz, Z., & DeDeo, S. (2020). From probability to consilience: How explanatory values implement Bayesian reasoning. *Trends in Cognitive Sciences*, *24*(12), 981–993.

Wolff, P., Barbey, A. K., & Hausknecht, M. (2010). For want of a nail: How absences cause events. *Journal of Experimental Psychology: General*, *139*(2), 191.

Woodward, J. (2010). Causation in biology: Stability, specificity, and the choice of levels of explanation. *Biology & Philosophy*, *25*, 287–318.

Zemla, J. C., Sloman, S. A., Bechlivanidis, C., & Lagnado, D. A. (2023). Not so simple! Causal mechanisms increase preference for complex explanations. *Cognition*, *239*, 105551. https://doi.org/10.1016/j.cognition.2023.105551

Zemla, J. C., Sloman, S., Bechlivanidis, C., & Lagnado, D. A. (2017). Evaluating everyday explanations. *Psychonomic Bulletin & Review*, *24*(5), 1488–1500.