Inference and selection in causal explanation

Stephanie Droop[1†], Tadeg Quillien[2], and Neil R. Bramley[2]

[1] Institute for Language

Cognition and Computation

School of Informatics

University of Edinburgh

Scotland

UK

[2] Psychology Department

University of Edinburgh

Scotland

UK

Author Note

Stephanie Droop https://orcid.org/0009-0007-6839-1406

† Correspondence concerning this article should be addressed to Stephanie Droop, stephanie.droop@ed.ac.uk.

Stephanie Droop was partially supported by UK Research and Innovation through the Center for Doctoral Training in Natural Language Processing and partially supported through a Turing AI Fellowship award, EPSRC grant ref - EP/V025708/1.

The authors declare that there is no conflict of interest regarding the publication of this article.

We confirm that we have followed ethical guidelines from the British Psychological Society Code of Conduct, and that this study sought and was given ethical approval by the University of Edinburgh Institutional Review Board (ref 693166).

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Stephanie Droop. The computational framework was developed by all authors, based on an initial draft by Tadeg Quillien. The overall empirical and computational strategy was guided by Neil Bramley. The first draft of the manuscript was written by Stephanie Droop and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data and code are available at our repository.

Acknowledgements: the authors thank Christopher Lucas for comments on an earlier draft of the manuscript.

Inference and selection in causal explanation

Explaining why an event occurred involves solving several related information-processing problems: One is inferring what may have actually happened (diagnostic causal inference). Another is highlighting which cause(s) (whether observed or inferred) contributed most significantly to the event's occurrence (causal selection). While much research has investigated causal inference and causal selection separately, little has studied them jointly. We report results of an experiment (N=215) examining how these aspects of causal reasoning interact, as is the case in real-world causal explanation settings. We study causal explanations in a forced choice task where participants choose from a set of singular variable-state explanations, in scenarios where the states of some variables are unobserved and may be partially or completely inferrable from the outcome. We show, with the help of computational modeling, that participants' aggregate behavior is consistent with the idea that they engaged in both inference and selection. Specifically, explanation-choice models that infer unobserved variables and select causes that covary with the outcome across counterfactual possibilities better predicted participant behavior overall than variants lacking either of these components. Moreover, exploratory modeling suggested that participants preferentially selected inferred over observed components to highlight in their explanations.

Keywords: causality; counterfactuals; explanation; inference

Introduction

Why did this car accident happen? Why did the dinosaurs go extinct? The drive to explain why particular events occurred is a core element of human psychology, and the basis of myriad inquests and official inquiries. In the field of causal cognition, the question of what intuitions underlie our choice of what to highlight in *causal explanations* of events has received a large amount of attention (Gerstenberg et al., 2021; Hitchcock & Knobe, 2009; Icard et al., 2017; Lagnado, 2021; Lombrozo, 2006; Lombrozo & Vasilyeva, 2017; Phillips et al., 2015; Quillien, 2020; Quillien & Barlev, 2022; Quillien & Lucas, 2023; Woodward, 2021).

Providing a singular causal explanation typically involves solving several related information-processing problems. In this paper we focus on two of the most important and examine how they interact:

- **Causal inference.** Causal inference is the process of figuring out what happened on the basis of the available evidence. For example, given a car crash (the effect in need of explanation) occurred while the driver was coming back from a party, how likely is it they were drunk (one potential cause)? Given it was winter in Canada, how likely is it there was ice on the road (another potential cause)?

- **Causal selection.** Causal selection consists in highlighting one or a few out of the several causes that contributed to an outcome (Hesslow, 1988; Quillien & Lucas, 2023). Suppose we believe both that the driver was drunk, and that there was ice on the road, and that both contributed to the accident. Which of these causes should we spotlight as *the* cause of this accident?

Both problems are crucial to causal explanation in everyday settings. Because the details of what happened leading up to an event are rarely completely transparent, a person looking for an explanation usually needs to piece together the available evidence and infer the states of various latent or unobserved factors. Since, in the real world, any given outcome is the end result of a complex interaction of many variables, people must also be selective to avoid producing

unhelpfully detailed explanations. In the causal cognition literature, questions of inference and of selection have almost exclusively been studied separately (with a few exceptions mentioned later). In this paper we study how people give causal explanations when they have to jointly solve both problems.

Inference

Causal inference concerns how people combine observations, interventions and knowledge of causal relations to make informed judgments about unobserved variables. These inferences may be general or specific, predictive or diagnostic, hypothetical or counterfactual. Causal inference and causal reasoning in humans is often studied with the help of the computational formalism of causal graphical models (Pearl, 2009). This general approach to understanding distinctively-human thinking patterns — by relating them to the representations and operations afforded by causal graphical models — is known as causal model theory (Holyoak & Cheng, 2011). Many experiments have found that people make inferences in ways that broadly approximate the normative prescriptions of causal model theory (Gopnik et al., 2007; Hagmayer et al., 2007; Lagnado, 2021; Meder & Mayrhofer, 2017; Sloman & Lagnado, 2004), or reflect plausible process-level considerations related to the computations it involves (Bramley et al., 2017; Davis & Rehder, 2020; Fernbach & Sloman, 2009; Sloman, 2022).

Causal inference is often divided into *type*- and *token*-level. *Type*-level inferences (also called causal discovery or structure learning) are about establishing the structure and parameters of a causal model that best describes a particular domain or repeatable situation (cf. Bramley et al., 2015; Coenen et al., 2015; Janzing & Schölkopf, 2010). The target of type-level inferences is identifying whatever is causally invariant in a context, and therefore able to support inferences at the token level. *Token*-level causal inferences leverage the reasoner's established causal model, combining it with observations or interventions, to draw inferences, or make predictions about the values of unobserved variables. Whereas type inference is concerned with relationships and rules, token inference concerns what states are taken by variables on a given occasion conditional on those relationships and rules. Often these states are things we can characterize as *events* (e.g., a

car crash occurring, or not). Studying token-level causal inference involves exploring how people make inferences about whether an event happened (or what state a variable took), on the basis of information about other events that happened (or other variables' states) (e.g., Hitchcock, 2007).

A classic setting for token-level inference is diagnosis: inferring the state of a potential cause (e.g., whether someone has a disease), based on observing occurrence of an event that might be its effect (e.g., because they have a certain symptom). For example, suppose event C often causes event E. We observe that E, and we want to infer the probability of C. We can solve this problem by using Bayes' rule to invert our causal model of C's influence on E:

$$P(C|E) = \frac{P(E|C)P(C)}{P(E)} \tag{1}$$

where the likelihood function $P(E|C)$, prior $P(C)$ and evidence $P(E)$ are drawn from the causal model describing the causal system.

Causal selection

In contrast to research on causal inference, research on causal selection investigates how people judge which of the factors that contributed to an outcome is the most important cause (Hesslow, 1988). For example, although the driver being drunk and the roads being icy both contributed to the accident, it is very often icy in winter in Canada, but less normal (and illegal) to drive drunk, making the drunk driving an intuitively better explanation as *the* cause of this accident.

The example just given is an intuitive illustration of a phenomenon that human causal selection typically exhibits, called *abnormal inflation*, whereby people attribute greater causal responsibility to causes that are rare, infrequent or otherwise abnormal, in situations where multiple causes are necessary, aka *conjunctive* situations (Gerstenberg & Icard, 2020; Icard et al., 2017). Conversely, when any cause is sufficient, aka *disjunctive* situations, people often show the inverted pattern, tending to highlight the success of the normal factor as most responsible for the outcome (Icard et al., 2017).

For simplicity, extant research on causal selection has often focused on experimental

settings where all variables are observed, so the reasoner already knows exactly what happened. Because of this, much less is known about causal selection in contexts where people also need to make inferences about what causes happened (what values all the cause variables took).

According to a popular family of accounts, people engage in causal selection by imagining other ways events could have turned out: counterfactual possibilities (Henne et al., 2019; Icard et al., 2017; Quillien, 2020), see also Gerstenberg et al. (2021). A recent computational model of causal selection based on this idea, the *Counterfactual Effect Size Model* (CES; Quillien, 2020; Quillien & Lucas, 2023) added to counterfactual simulation the idea of an *effect size*. The model holds that people judge whether event *C* was a cause of event *E* by: i) simulating many different alternative ways the situation could have happened, and ii) computing a measure of the dependence between *C* and *E* across these imagined possibilities. The more *E* counterfactually depends on *C*, the better candidate it is as a causal explanation. The measure of counterfactual dependence computed by the model is the Pearson correlation coefficient between *C* and *E* across the simulated counterfactual worlds (or, in cases where there is confounding between C and E, an appropriately corrected measure). Note that other measures of causal effect familiar from the causal modeling literature, like the Average Treatment Effect or Pearl's Probability of Necessity and Sufficiency (Pearl, 2009), do not account for people's causal selection judgments as well as the CES measure (see Morris et al., 2019; Quillien, 2020; Quillien & Barlev, 2022).

The CES model can parsimoniously explain data from past experiments on causal judgments (Gerstenberg & Icard, 2020; Icard et al., 2017; Lagnado et al., 2013; Morris et al., 2019). It has also made successful new predictions, both in simple experimental settings (Konuk et al., 2026; O'Neill et al., 2025; Quillien & Lucas, 2023) and in a real-world context (Quillien & Barlev, 2022). Although the model fits people's judgments well in fully observed settings, it has not been tested in settings where the states of some explanatory variables are uncertain.

Actual causation

Actual causation is a token-level categorical judgment as to whether one thing contributed at all to the occurrence of another (Halpern, 2016). Generally questions of causal selection are

restricted to things that are deemed to be *actual causes* of the effect. As we will be modeling explanations that refer to different states of variables, it becomes important to consider the role of these categorical actual causation judgments as well as selection values. For example, there may be ice on the road, but if we knew that a drunk driver crashed before reaching it, then we know this ice was not an actual cause of this accident despite it having a type-level causal relationship with car accidents in general. A universal criterion for actual causation has proven hard to formalize (Baumgartner & Glynn, 2013); however several technical accounts exist (Halpern & Pearl, 2005a; Halpern, 2016).

Blending inference with selection and explanation

Although inference and selection have historically been studied separately, recently some authors have studied scenarios where they meet. One study, Pacer et al. (2013), asked participants to select or rate explanations in a context where they knew the relevant causal model but observed only the outcome variable. They found that participants' explanations did not simply track the posterior probability of the relevant variables' states given the effect, nor did participants simply pick states that would maximize likelihood of observing the effect; heuristic models which set the variables progressively could capture participant behavior better (Nielsen et al., 2012; Yuan et al., 2011). While these heuristic models hinted at counterfactual considerations, through the idea of causal information flow (Nielsen et al., 2012), these experiments did not provide a way to distinguish between inference and selection as distinct computational problems. One setting where inferences feed into selections is found in Gerstenberg et al. (2018). This paper found that modeling how people update their beliefs is key to accounting for blame and praise judgments in some social scenarios, and that these judgments also involve counterfactual reasoning. Ying et al. (2025) similarly developed an account of how people attribute mental states to explain behavior. Their account holds that people attribute a belief to an agent when both the posterior probability that the agent holds this belief is high, and the belief is counterfactually relevant for the observed behavior.

Other research has examined inferences *from* selection explanations. These inferences

generally work by inverting the logic of causal selection (Kirfel et al., 2022; Nam et al., 2023; Navarre et al., 2024). For instance, Kirfel et al. (2022) found that when people are given an explanation for why something happened, they can use this explanation to infer an event's normality if they know the causal structure, or infer the causal structure if they know the events' normality. Nam et al. (2023) found that people could infer causal structures better when observations were paired with selection judgments, and Navarre et al. (2024) similarly found that people could better infer the functional form of a relationship between multiple causes and an effect, when given selection explanations as well as observations. None of the above projects, however, has explored the mixed setting where both observed and unobserved variables are potential candidates for selection judgments.

Explanations as sharing computations?

This seems like a salient omission in part because explanation, as a cognitive phenomenon, is often related to the problems of compressing and packaging information both for storage and communication (Kinney & Lombrozo, 2024; Harding et al., 2025; Quillien et al., 2025). Explanations are also fundamentally "three-place predicates": [*someone*] explains [*something*] to [*someone*] (Hilton, 1990). In particular, explanations might focus on informing the explainee of something that they did not already know (Halpern & Pearl, 2005b). The selective nature of explanations is advantageous in part because it provides such information: mentioning that C was the cause of E might be useful if the explainee did not already know that C happened, but it can also inform the explainee about the prior probability of C, or the causal structure of the system (Kirfel et al., 2022; Nam et al., 2023; Navarre et al., 2024). Explanation can have an informative function even in contexts where sender and receiver have access to the same information. Think of two people working on a puzzle: the one who generates an insight first can share it with their partner, provisioning them with the outcome of their successful cognitive work, rather than with information they fundamentally lacked. Accordingly, we will use our analysis to explore whether people treat unobserved (wholly or partly inferrable) variables differently from directly observed variables in their explanations.

Extended and collective cognition depends on distributing computational "work" and communicating its products (Christian & Griffiths, 2016; Dijkstra, 2002). The act of provisioning someone with the products of computational work is sometimes known as "computational kindness" (Christian & Griffiths, 2016). In this spirit, we speculated that in situations of uncertainty, people might favor explanations that highlight events on whose occurrence the effect *not only* counterfactually depends, but that *also* had to be inferred from that observation. That is, if there are two options that are equally explanatory but one is directly observable and the other had to be inferred, it seems plausible that a helpful explainer would prioritize mention of the latter over the former.

The current work

In this paper, we study causal explanations in a context where some events are observed and some are unobserved. For example, in one scenario, a software enterprise is giving a product demonstration to a client and will win a contract if they present either of two new features, provided the files underlying the features do not malfunction. Suppose we observe that the sales engineer presented both features and won the contract, and also know how often they present each feature and how often the files work, but we don't know which files actually worked and so which feature(s) actually won the contract this time. Why do participants think the contract was won? This task invites inference (because two of the causal variables are unobserved and their states can be inferred from the observation) as well as causal selection (because there are multiple candidate causes they could point at).

We outline a computational framework for causal explanation in the presence of unobserved variables, and report results of an experiment unpacking participant behavior through the lens of this model.

Computational framework

Our computational model of this setting involves several somewhat-separable components: Inference; Selection; as well as considerations of Actual Causation and Computational Kindness. We introduce these components first as a general mathematical

framework, and then apply the framework to our specific experimental scenario. The modularity of our modeling approach allows us to consider a nested set of "lesioned" model variants, providing a means to assess the relative contribution of components in accounting for participants' response preferences. Our models all assign a score $\mathcal{C}(X = x)$ to each (potential) causal event, such that an event with a higher causal score is a better candidate, and so a more likely choice, as a causal explanation. To outline, the calculation of this causal score comprises:

1. **Inference:** In the process of making any explanation, a normative learner should update the probability of unobserved variables by conditioning on observed variables using Bayes theorem.

2. **Causal selection:** They may also assess the causal responsibility of each variable state (under each possible state of affairs) as they do in fully observed settings. We used the *Counterfactual Effect Size* model (CES; Quillien & Lucas, 2023) to measure the effect's dependence on each potential variable state and convert this into a selection score, but other selection models are possible (e.g., Icard et al., 2017; Lagnado et al., 2013).[1]

3. **Actual causation:** A learner ought to restrict their possible choices in a given context to those variable states that have a valid causal pathway to influence the effect.

4. **Computational kindness:** A learner may also privilege explanatory variables whose probabilities changed upon observing the effect (assuming they performed inferences), or are in some other way useful things to highlight. Whether a variable was observed or inferred may therefore contribute to final choice scores.[2]

---

[1] We acknowledge that the word 'selection' can be ambiguous between the theoretical construct of 'selection' discussed here, and the broader behavioral task of 'selecting' a variable to supply as the singular causal explanation. To minimize confusion throughout this paper we use the word 'selection' for the former meaning and 'choice' for the latter.

[2] Our inclusion of this component is exploratory, since our experiment was not designed to test it, and there are multiple plausible ways to operationalize it. We compare several options in Appendix D.

General mathematical framework

Here we give a general formalization of our proposal for how inference and selection could be combined, before illustrating how this proposal applies to our particular experimental situation.

We assume the reasoner knows the causal structure of the relevant system. We operationalize this structure as a Structural Causal Model (SCMs; Pearl, 2009). Under a slight simplification of SCM formalism (see Ibeling & Icard, 2023, for details), an SCM is a tuple $M = \langle \mathcal{U}, \mathcal{V}, \mathcal{F}, P(\mathcal{U}) \rangle$, where $\mathcal{V}$ is a finite set of endogenous variables, $\mathcal{U}$ is a finite set of exogenous variables, $\mathcal{F}$ are the structural functions connecting them, and $P(\mathcal{U})$ is a probability distribution on joint valuations of $\mathcal{U}$. In practice this means exogenous (root) variables $\mathcal{U}$ come from probability distributions, while endogenous variables $\mathcal{V}$ (i.e., variables with parents) are functions of their parents, i.e., *structural equations* describe the causal relationships between the variables. The causal relationships among variables $\mathcal{V}$ are thus deterministic, and stochasticity in the system comes from $\mathcal{U}$, and subjectively from incomplete observation of the variables. Binary variables labeled with capital letters represent whether each event occurs (for example $X = 1$ means that event X happened and $X = 0$ means it did not happen).

We consider a structure where the effect of a cause component X on effect E depends on an unobserved (latent) noise variable X_u which acts as its causal strength (Cheng, 1997). X is in $\mathcal{V}$ and X_u is in $\mathcal{U}$. X is itself caused by exogenous variables not shown. For a single cause X and effect E depending on X the structural equation is:

$$E := (X \wedge X_u) \tag{2}$$

In other words, X_u determines whether $X = 1$ causes $E = 1$ and thus E occurs only if both X and X_u occur; see Figure 1 for a simplified schema of one node pair. Functionally what this means for our setup is that all cause variables occur on a given occasion according to some known prior probability distribution. The important difference is that we can observe whether the variables in $\mathcal{V}$ actually occur but we cannot do this for the variables in $\mathcal{U}$. These pairs of causes

— an observed and an unobserved counterpart — can then be composed into larger structures.

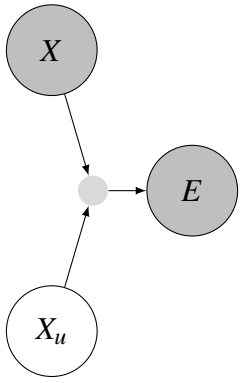


Figure 1

A single node pair of variables: X is drawn from $\mathcal{V}$ and X_u is drawn from $\mathcal{U}$. X is defined by exogenous variables not shown. The small gray joining node simply illustrates that X and X_u form a conjunctive pair. (Large) gray nodes are observed; white nodes are unobserved. The occurrence of X and X_u are (unconditionally) independent.

To give a causal explanation for the state of E that happened, the reasoner should i) infer the posterior probability of the latent variables given what they have observed; ii) engage in causal selection; iii) integrate the two processes.

We consider whether variable realization $X = x$ was the cause of outcome $E = e$. We define $S(X = x)$ as the selection score of event $X = x$. The challenge is to develop an extension of causal selection that delivers causal scores even in the presence of uncertainty about the values of X_u. For convenience we will denote the posterior distribution in abbreviated form as $P(X_u|X,E) = P_\alpha(X_u)$. We denote the prior distribution as $P(X)$.

We denote **W** the set of variables other than E and X. We will also write $S(X = x|\mathbf{W{=}w})$ to express the selection score that would be assigned to $X = x$ under the assumption that **W=w** in the actual world (this is useful notation when we need to consider several possible hypotheses about the actual world consistent with our observations).

Causal inference

On the basis of an observation we infer a posterior over unobserved variables using Bayes' rule, and the likelihoods defined by the causal model.

Causal selection through counterfactual simulation

The CES model holds that people engage in causal selection by simulating counterfactual possibilities. Specifically, the model samples counterfactual worlds from the structural causal model describing the relevant causal system, starting with an assumed true state for each variable, and uses these counterfactual worlds to compute a measure of causal 'effect size' quantifying how much the event $X = x$ caused the effect $E = e$.

Each counterfactual world is simulated by sampling each variable from a probability distribution, and then setting the effect variables according to their structural equations. When simulating a possible counterfactual world, each variable Y that has no parents in the causal model is sampled from the probability distribution $s\delta(Y) + (1-s)P(Y)$, where $\delta(Y)$ is a probability distribution that assigns 1 to the value that Y takes in the actual world and 0 otherwise, $P(Y)$ is the prior probability of Y, and s is a 'stability' parameter governing the degree of anchoring to what happened in the actual world (Lucas & Kemp, 2015; Quillien et al., 2023).[3]

The CES score of $X = x$ for $E = e$ is then computed on the basis of the simulated possibilities. This score is an 'effect size' that computes to what extent X has an effect on E across the simulated possibilities. In our setting, the effect size computed by the CES model is the Pearson correlation coefficient between $X = x$ and $E = e$ across simulated worlds (Quillien, 2020). This correlation coefficient has a causal interpretation because in our setting there is no confounding between X and E—a corrected formula would be used in confounded settings (see Quillien, 2020; Quillien & Lucas, 2023).

We denote the CES score for $X = x$ as $S(X = x)$. We also normalize the CES scores by passing them to a softmax transformation, so that the CES score of each variable is expressed

[3] For a generalization of this way of defining a distribution over possible counterfactual worlds, see for example Von Kügelgen et al. (2023).

relative to the score of the other candidate causes, yielding a score $\hat{S}(X = x)$:

$$\hat{S}(X = x) = \frac{\exp(S(X = x)/\tau_1)}{\sum_{X_i \in \vec{X}} \exp(S(X_i = x_i)/\tau_1)} \tag{3}$$

where $\vec{X}$ is the set of candidate causes of $E = e$, and x_i is the value of candidate cause X_i in the actual world. We perform this transformation following evidence that people might normalize their causal selection judgments in this way (Morris et al., 2019). This step also allows the model to modulate the extent to which counterfactual simulation affects causal judgment, via the temperature parameter τ_1 (high values of the parameter translate into a weak effect of counterfactual simulation on the overall causal judgment relative to other factors such as posterior probability).

Applying the CES under uncertainty

The CES model is defined for fully observed settings. To apply it to situations with unobserved variables (where this assumption doesn't hold), we must make some choices as to how to handle the uncertainty over $\mathcal{U}$. One intuitive way to do this is to compute a selection score (here a CES score) for each possible state of the actual world compatible with what we know, and then compute a weighted average of these scores, where the weights are the posterior probabilities of the states of the world.

Formally, we compute a marginal causal selection score $S'(X = x)$ by i) assuming that $X = x$ in the actual world, and ii) marginalizing across all possible values of the other variables, weighted by their posterior probabilities:

$$S'(X = x) = \sum_{\mathbf{w}} \hat{S}(X = x | \mathbf{W} = \mathbf{w}, X = x) P_\alpha(\mathbf{w} | X = x) \tag{4}$$

Computing the marginal causal score for unobserved variables introduces an additional complication: we may also not be able to infer definitively whether the variable really took the state we are calculating the explanatory value of. One intuition is that people will tend to say '$X_u = 1$ caused the outcome' if i) it is in fact likely that $X_u = 1$ in the actual world, and ii) $X_u = 1$ has a high S' selection score in that contingency. One way to implement this is to compute an

expected counterfactual effect size $\tilde{S}$ by multiplying the marginalized CES propensity score S' by the posterior probability of the variable value, $X = x$.

$$\tilde{S}(X = x) = S'(X = x)P_{\alpha}(X = x) \tag{5}$$

Valuing "computational kindness": Privileging inferred variables

We hypothesized that explainers might favor explanations that mention variables whose state they had to infer, all else being equal. Since an explainer must perform inference to obtain posterior probabilities of unobserved variables, we will consider a straightforward descriptive operationalization of this idea by considering models that add a constant bonus to all inferred variable states (here, unobserved variable states whose posterior probability is positive). Formally, we give a constant bonus $\mathrm{K}_{\alpha}(X) = 1$ to all choices that neither refer to observed variables or inferred-as-impossible variable states. This bonus is scaled by a free parameter κ; see Equation 6.

We also compare this simple operationalization with several more complex, information-theoretic measures, including prior-to-posterior Kullback-Leibler divergence, Information Gain, and the Mutual Information between unobserved and target variables; see Appendix D.

Overall causal score

Finally, we compute the overall causal score $\mathcal{C}(X = x)$ by taking a linear combination of the expected counterfactual score $\tilde{S}$ and the computational kindness K_{α}, and checking for actual causation:

$$\mathcal{C}(X = x) = (\tilde{S}(X = x) + \kappa \mathrm{K}_{\alpha}(X))\mathbb{I}_{\mathrm{AC}}[X = x] \tag{6}$$

where $\mathbb{I}_{\mathrm{AC}}[X = x]$ is 1 if $X = x$ is an actual cause of E, and 0 otherwise.

Singular causal explanation with two node pairs

We now illustrate how this model works for our specific case, while explaining the rationale for some of our modeling choices.

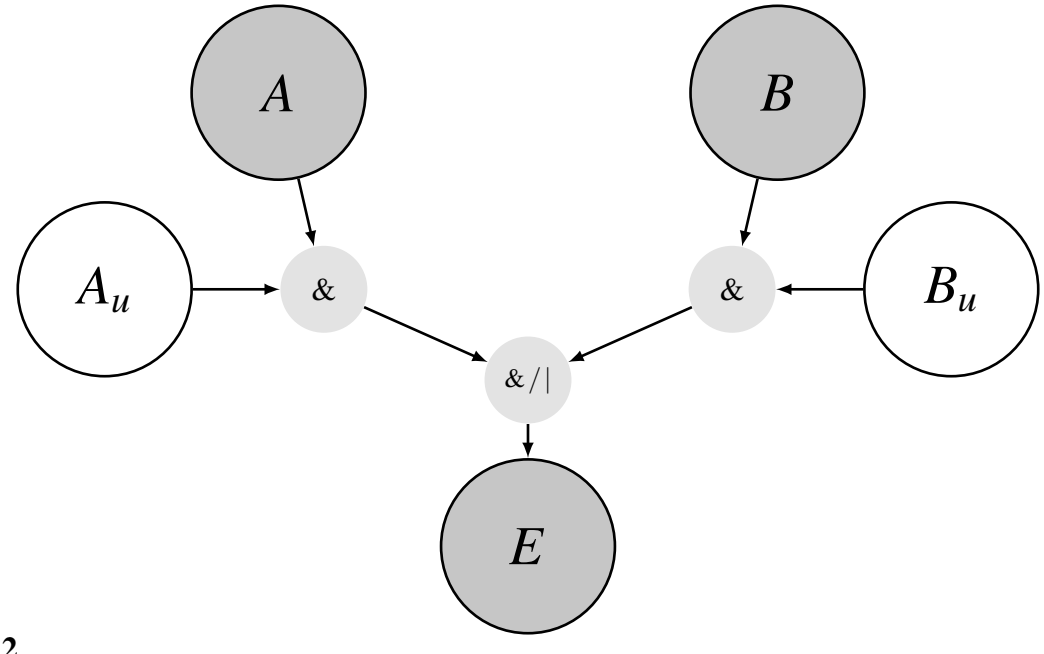


Figure 2

Graphical representation of our causal structure. Grey nodes (A, B, E) denote observed variables; white nodes (A_u, B_u) denote unobserved variables.

We consider causal systems where two observable variables A and B causally influence outcome variable E. Each has an unobserved counterpart, A_u and B_u, which determine whether A and B respectively can have an effect on E. Here A, A_u, B, B_u are mutually independent. A and B are generated by classic exogenous variables not shown. Figure 2 shows a graphical model of such a causal system. We study a conjunctive and a disjunctive structure. In the conjunctive structure E only happens if all variables happen:

$$E := (A \wedge A_u) \wedge (B \wedge B_u) \tag{7}$$

In the disjunctive structure, E happens if either A and A_u happen, or if B and B_u happen:

$$E := (A \wedge A_u) \vee (B \wedge B_u) \tag{8}$$

While the values of A and B are observed, the values of A_u and B_u are not. To give a causal explanation for the state of E that occurred, the reasoner should i) check what events *could* contribute to the outcome (actual causation) ii) engage in causal selection, iii) infer the posterior

probability of the latent variables given what they have observed: $P(A_u, B_u|A, B, E)$, iv) integrate these processes. We discuss each component in turn.

Actual causation

In our computational modeling we used a heuristic method to exclude events that do not qualify as actual causes. Specifically, we assigned a causal score of $\mathcal{C}(X) = 0$ to any variable state $X = x$ that does not match the value of the outcome (e.g., if $E = 1$, then $B = 0$ is not an actual cause of E). This is a simplified version of the first of three technical conditions in Halpern & Pearl (2005a). We also assigned a causal score of $\mathcal{C}(X) = 0$ to unobserved variables if their conjunctive counterpart has value 0. For example if E happens and $B = 0$ and $B_u = 1$, then intuitively neither $B = 0$ nor $B_u = 1$ was a cause of E. This heuristic works here because the causal functions are never preventative. For more sophisticated computational accounts of determining categorical actual causation see, for example, Halpern & Pearl (2005a); Halpern (2016).

Causal selection

We assume the explainer computes a causal selection score for each setting of all variables. For our purposes we use the CES as outlined under the 'Causal selection through counterfactual simulation' heading. Concretely, we sampled 100,000 counterfactuals for each world assuming a stability of $s = .7$ on the basis of past model fitting (Lucas & Kemp, 2015; Quillien & Lucas, 2023).

We compute the CES scores for each variable, S, under all combinations of unobserved variables states. For each combination, we apply a softmax over the causal selection scores of the four variables, with temperature τ_1, to yield a selection propensity score, $\hat{S}$. For example, having observed A, B and E occur in a disjunctive situation, one possibility is that A_u and B_u both *also* occurred. Therefore, for this contingency we calculate the CES scores for $A = 1$, $B = 1$, $A_u = 1$ and $B_u = 1$ and convert them into selection propensities:

$$\hat{S}(x) = \frac{\exp(\mathcal{S}(x)/\tau_1)}{\sum_{x \in \{A, A_u, B, B_u\}} \exp(\mathcal{S}(x)/\tau_1)}. \tag{9}$$

Causal inference

We assume the explainer infers the posterior distribution over A_u and B_u using Bayes' rule:

$$P(A_u, B_u | A, B, E) = \frac{P(E|A_u, B_u, A, B) P(A_u, B_u)}{P(E|A, B)} \tag{10}$$

For convenience we will denote the posterior distribution in abbreviated form as $P(A_u, B_u | A, B, E) = P_\alpha(A_u, B_u)$.

Integrating inference and selection

Finally, to arrive at marginal selection scores, our explainer computes a weighted average over the possible $\hat{S}$ values, under the possible settings of unobserved variables A_u and B_u, using a specific form of the general Equation 4. For example, the marginal selection score for $A = a$, denoted $S'(A = a)$ is:

$$S'(A = a) = \sum_{A_u, B_u} \hat{S}(A = a | A = a, B = b, A_u, B_u) P_\alpha(A_u, B_u) \tag{11}$$

where a and b are the actual-world values of A and B, and $\hat{S}$ is the selection propensity score for each variable in each hypothetical combination of variables from the earlier step, $\hat{S}$.

Finally, we arrive at an *expected* counterfactual effect size $\tilde{S}$ by multiplying the marginal selection score S' by the posterior probability of the variable taking that value as in Equation 5. For example for $A_u = 1$:

$$\tilde{S}(A_u = 1) = S'(A_u = 1) P_\alpha(A_u = 1) \tag{12}$$

Combining the cognitive modules

We assume the various elements of inference, computational kindness, actual causation and selection are combined as in Equation 6 to produce a causal score $\mathcal{C}(x)$ for each candidate explanation.

Choice model

The sections above specify how the model assigns causal scores to variables. To convert these causal scores to predicted choice proportions over the possible categorical choices, we assume that participants mix occasional random choices with choices that softmax over the causal scores:

$$P(\text{choice} = x) = \frac{\varepsilon}{|\mathcal{X}|} + (1-\varepsilon)\frac{\exp(\mathcal{C}(x)/\tau_2)}{\sum_{x\in\mathcal{X}}\exp(\mathcal{C}(x)/\tau_2)} \tag{13}$$

$\mathcal{C}(X)$ is the composite causal score. $\mathcal{X}$ is the set of variable states participants select from $\{A=0; A=1; A_u=0; A_u=1; B=0; B=1; B_u=0; B_u=1\}$, τ_2 is a temperature parameter (higher values indicating more stochasticity) and ε is a random responding parameter, assigning some probability mass to responses that otherwise have no support under that model variant. These responses always included whatever values observed variables A and B did not take.

Lesioned models

We also fit 'lesioned' models to assess our hypotheses. Our primary hypothesis was that when people make a causal judgment, they engage both in inference (about the value of unobserved causes) and in causal selection. Considering model variants with one or other component missing we can test this hypothesis directly. Because we also explored the role of actual causation and computational kindness, we also consider variants with and without these elements. This gives four 'modules' which can be switched in and out relatively independently (except that the kindness module only contributes if inference is performed). We hypothesized all modules would jointly help account for how people provide causal explanations which would be evidenced by our full model outperforming all lesioned variants after accounting for their differences in flexibility. Thus, the full model contains all four modules, with $C(X)$ as above, and we progressively lesion by removing modules from $C(X)$ as follows.

Lesioning actual causation

We test variants of the models defined above that *do not* check if an event is an actual cause of the outcome ($\mathbb{I}_{\text{AC}}[]$ is fixed to 1 in Equation 6). In practice, this means causal scores are calculated for all variables in all situations, even when the context prohibits them from causally influencing the outcome.

Lesioning causal selection

To lesion the causal selection module, we assume that people do not engage in counterfactual simulation. Once they determine which variables are actual causes of E, they consider all to be equally good (CES scores in Equation 9 are always 1). Note this does not result in uniform final selection predictions since, other things being equal, unobserved variable states are less likely to be selected than observed variables due to their lower posterior probabilities (but also could be *more likely* to be selected, due to computational kindness considerations).

Lesioning inference

To remove the inference module, instead of setting $P_{\alpha}(A_u, B_u)$ to the posterior, we 'freeze' it as the prior distribution. That is, we have $P_{\alpha}(A_u, B_u) = P(A_u, B_u)$. The model otherwise works as above.

Lesioning both inference and selection

This model assumes that people select among actual causes almost indiscriminately. That is, they assign a causal score of $\mathcal{C} = 1$ to observed variable values, and a causal score of $\mathcal{C} = P(X_u = x_u)$ to unobserved variables, where $P(X_u = x_u)$ is the prior probability of that variable.

Lesioning kindness

We also consider all with-inference models with and without our exploratory additional kindness component, which increases the probability of selecting an unobserved variable to degree κ. When this is lesioned, κ is set to 0, and otherwise the series of models works as above.

Experiment

Methods

We conducted a behavioral experiment to investigate how participants choose explanations under uncertainty.[4]

Design

Table 1

Observed scenarios: Variables A, B, E

	Conjunctive				Disjunctive		
	A	B	E		A	B	E
c1	0	0	0	d1	0	0	0
c2	0	1	0	d2	0	1	0
c3	1	0	0	d3	0	1	1
c4	1	1	0	d4	1	0	0
c5	1	1	1	d5	1	0	1
				d6	1	1	0
				d7	1	1	1

Table 2

Probability settings 1:3

Prior	Set1	Set2	Set3
$A = 1$	.1	.5	.1
$A_u = 1$	.5	.1	.7
$B = 1$	.8	.5	.8
$B_u = 1$	.5	.8	.5

Note: The four root variables are mutually independent.

Note. Instead of the word *condition*, certain terms have a specific usage from now on in this project:

- A *scenario* is one of the 12 possible combinations of observed variables A, B and effect E (Table 1).
- A *probability setting* is one of the 3 vectors of prior probabilities of the four variables A, A_u, B, B_u (Table 2).
- A *world* is one of the 36 combinations of scenario and probability setting engendered by these two tables.

We investigated how participants select causes in scenarios containing four binary causes: two observed variables A and B, and latent success rates A_u and B_u, where the variables are grouped in two pairs as in Figure 2. We asked each participant to select one of the eight possible

[4] You can try the experiment here (at comprehension quiz, select: Yes, No, True, 12).

variable states as a causal explanation for variable E's occurrence or non-occurrence. Within subject, participants make these selections for all of the 12 different logically possible combinations of observed variables and effect E ('scenarios', shown in Table 1[5] and Appendix A) for all possible settings of unobserved variables conditional on each observation type. Each trial presented the underlying causal structure with a vignette, including prior probabilities for all four variables, along with a simplified graphical representation of the requisite causal structure. Then participants were shown a concrete state of observed variables ('what happened *this time*'), and were asked to explain the outcome $E = 1$ or $E = 0$ by selecting one of the eight possible variable values ($4 \times \{0, 1\}$); Figure 3.

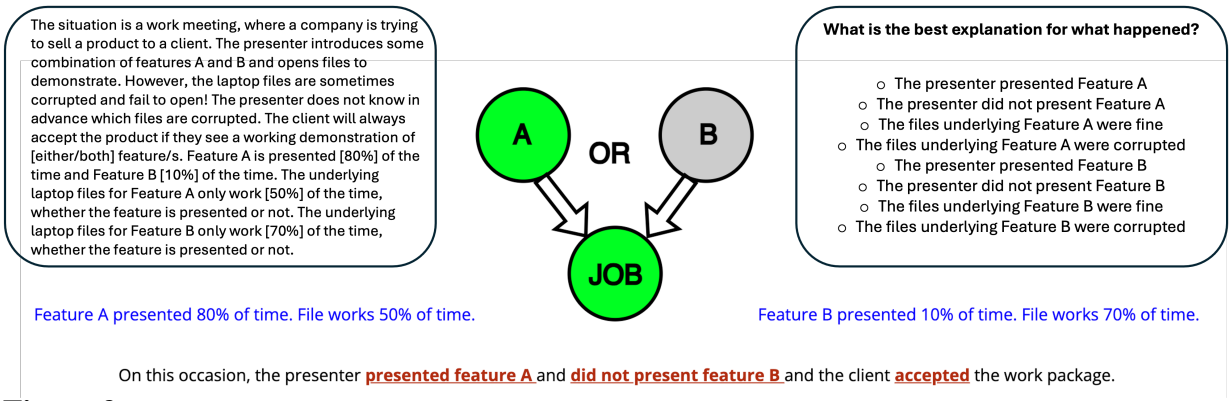


Figure 3

Simplified schematic of one trial: blue text gives base rates; grey/red text and graph describe what happened this time. A green node on the figure in the middle indicates variable is 1 and gray indicates 0, meaning here $A = 1, B = 0, E = 1$.

We used three cover stories: 1) entering a cafe, 2) a university reading group and 3) a client meeting. We manipulated the prior probabilities of the causally relevant variables independently to the cover story. Instead of varying the prior probabilities continuously at combinatorial expense, we chose three settings (see Table 2) to tease apart relationships where 1)

[5] Five of the 12 scenarios involve the conjunctive structure defined in Equation 7, and seven involve the disjunctive structure defined in Equation 8. Unequal split is due to the fact some observation patterns are possible for the disjunctive but not conjunctive structure (e.g., $A = 1$, $B = 0$, $E = 1$).

the two observed variables strongly vary in their normality and the unobserved variables are balanced; 2) the unobserved variables strongly vary and the observed are balanced; 3) all vary. For each trial, one of the three probability settings and one cover story was randomly selected. Thus, on average participants saw each cover story four times, and each probability setting four times.

For example, in the 'client meeting' cover story, participants read:

> The situation is a work meeting, where a company is trying to sell a product to a client. The presenter introduces some features in a talk, and then opens files to demonstrate. However, the laptop files are sometimes randomly corrupted and fail to open! The presenter does not know in advance which files are corrupted. The product has Feature A and Feature B, and the client will always accept the product if they see a working demonstration of [either/both] feature[/s]. The company presents Feature A [10%] of the time and Feature B [80%] of the time. The underlying laptop files for Feature A are sometimes corrupted: they only work [70%] of the time, whether the feature is presented or not. The underlying laptop files for Feature B are sometimes corrupted: they only work [50%] of the time, whether the feature is presented or not.

This is the same cover story shown in Figure 3. See Appendix B for the others.

Participants

We recruited 215 participants (111 female, 1 other, age Mean $\pm$ sd 38.5 $\pm$ 12.8, range 18-80, all fluent in English and fulfilling the criteria of being active users of the Prolific platform, defined as using the site within preceding 90 days and having a completed profile) using the Prolific subject pool. They were paid £3.69 and the experiment took Mean $\pm$ sd 31.5 $\pm$ 13.9 minutes. The sample size was chosen so that circa 72 saw each world. Informed consent was obtained from all individual participants included in the study. We followed ethical guidelines from the British Psychological Society Code of Conduct, and were given ethical approval by the University of Edinburgh Institutional Review Board (ref. 693166).

Stimuli

Each trial presented a series of text and pictures following a fixed format. Stimuli were created using JSPsych 6.3.1 html plugins (De Leeuw, 2015). The general schema presented the base rates at which all four events usually happen, and the causal setup (i.e. whether the scenario was conjunctive — both observable events and their latent enablers needed for the outcome to occur, or disjunctive — just one and its latent enabler). We then revealed the value of the observed variables *this time*. See Figure 3 for an example of the work meeting for the disjunctive scenario where $A = 1$, $B = 0$, $E = 1$. Finally participants selected one among all eight possible explanations: in the example shown in Figure 3, plausible explanations may include "The presenter presented Feature A" ($A = 1$), "The files underlying Feature A were not corrupted" ($A_u = 1$).

Procedure

The experiment was implemented in JavaScript. Participants were recruited from Prolific and completed the experiment in the browser on their own devices. After calibrating their computer screen, they were presented with the study's information sheet and consent form. Participants were then given instructions for completing the experiment and shown examples of the stimuli. They then completed a four-item quiz to test their understanding before beginning the experiment. All participants saw all 12 scenarios one-by-one in a random order, with a random allocation of one of the three probability settings. The left/right presentation position on screen of the variables and their prior probabilities was counterbalanced between participants.

Analysis

Data were analyzed using `R` 4.1. Package *lme4* (Bates et al., 2014) was used for mixed-effects regression models following recommendations of Meteyard & Davies (2020), via package *lmerTest* (Kuznetsova et al., 2017) for tests. The data and code for modeling and analysis are available in our Anonymized Repository.

Results

A Fisher's exact test found no statistical difference between the three cover stories, for each of the 36 worlds[6] and so our analyses collapse across cover stories.

Figure 4 shows the choice proportions of participants in probability setting 3 [$P(A) = .1, P(A_u) = .7, P(B) = .8, P(B_u) = .5$]. We chose to visualize setting 3 first in the main text because all the base rates of the four variables are different and so it is representative of the overall pattern, with qualitative conclusions similar across all conditions. A later qualitative analysis focuses on behavioral patterns in settings 1 and 2 which isolate probability effects stemming from the observed or unobserved variables respectively.

Participants' judgments were clearly non-uniform across the worlds (all item-level goodness-of-fit $\chi^2 > 50$, df = 7, $p < 4.1e-7^{***}$ Bonferroni-adjusted for all 36 worlds). This simply establishes that participants' responses depended systematically on the value of the variables they observed.

We tested whether participants' judgments depended on the particular probability setting. An omnibus χ^2 test of independence indicated some difference in distribution between the three probability settings, $\chi^2(186, N = 2580) = 238.8$, $p = .016^*$ after Bonferroni-adjustment for three comparisons, Cramer's $V = .215$.[7]

In the following subsections we first fit the full and all lesioned models in aggregate and by participant, and comment on the best fitting models and the relative contribution of each of the four modules of causal cognition. Then we illustrate these quantitative results by exploring qualitative patterns in participants' choices.

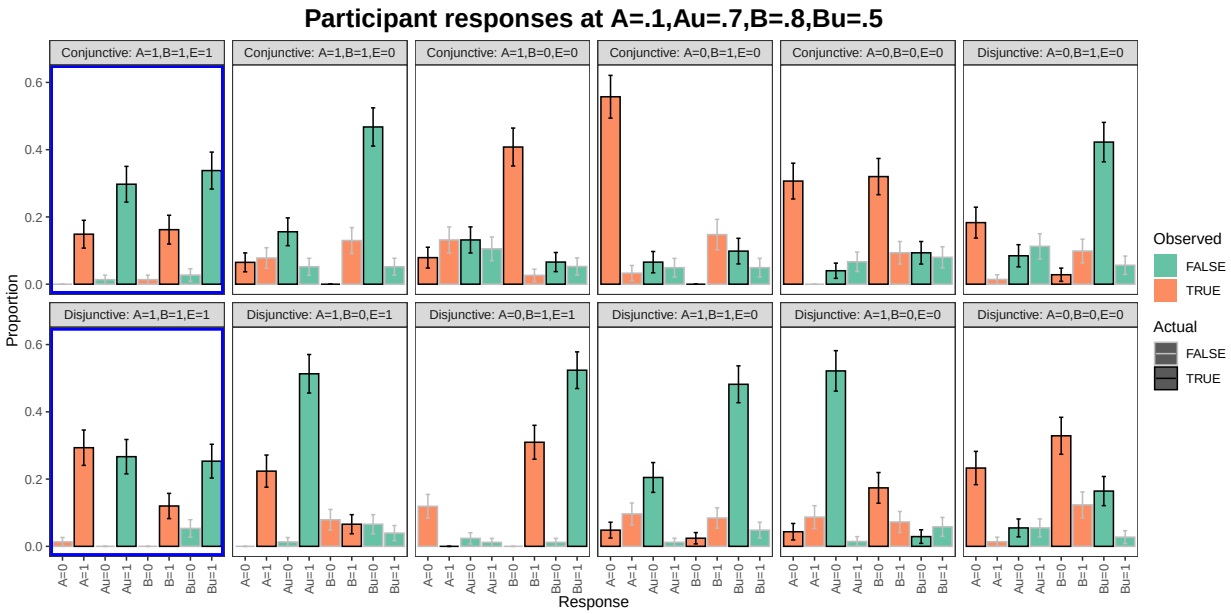


Figure 4

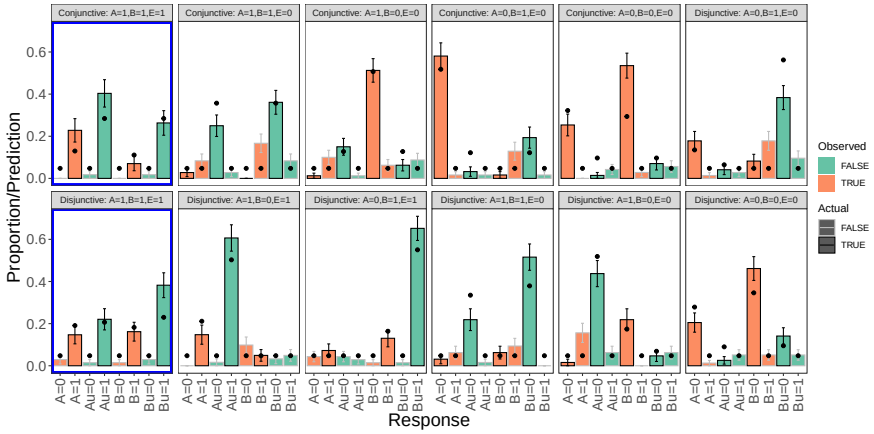
Results in Setting 3 $P(A) = .1, P(A_u) = .7, P(B) = .8, P(B_u) = .5$. *Proportion of participants selecting each explanation (bars). Blue highlights the canonical "everything happened" scenario* ($A = 1$, $B = 1$, $E = 1$). *Bars show proportion of responses each variable state received in each world; fill indicates whether the node was observed (orange) or unobserved (green). Error bars show* ± 1 *SE of the observed proportion, colored black for Actual causes and gray for non-Actual causes.*

Model fit

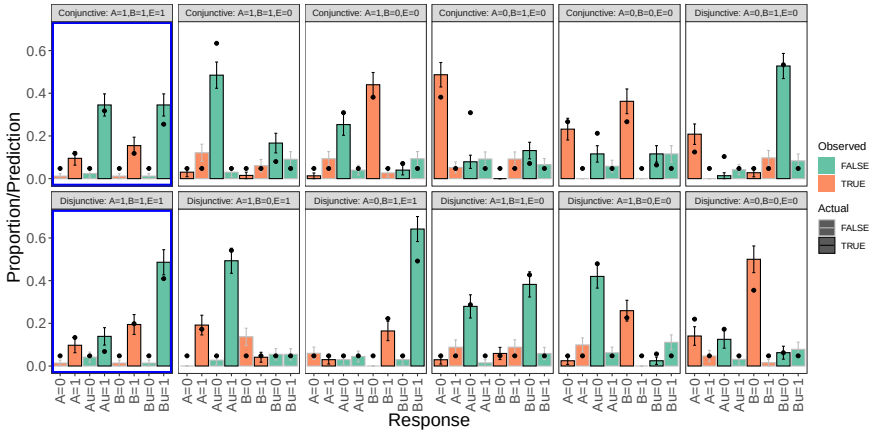
We fit the models to the data by minimizing negative log likelihood, with two softmax temperatures τ_1 (for models with selection) and τ_2, the noise ε, and the kindness weight κ (for models with `Kindness`) as free parameters, optimized with Brent method as implemented by R's `optim` function. Table 3 displays the model fits. The four modules of `Kindness`, `causal Selection`, `Actual Causation` and `Inference` were progressively lesioned, and absence of a

[6] A *world* is one of the 36 combinations of 12 observable scenarios of variables *A*, *B* and *E* and the 3 probability settings; see Table 1 for definitions.

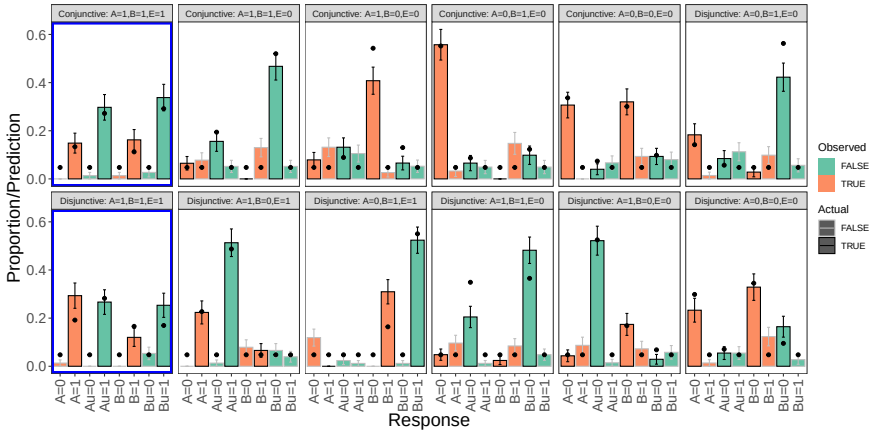
[7] Two choice-scenario combinations were never selected from the 96 (8 choices * 12 scenarios) so the test was over three vectors of length 94.



(a) *Setting 1* $P(A) = .1, P(A_u) = .5, P(B) = .8, P(B_u) = .5$.



(b) *Setting 2* $P(A) = .5, P(A_u) = .1, P(B) = .5, P(B_u) = .8$.



(c) *Setting 3* $P(A) = .1, P(A_u) = .7, P(B) = .8, P(B_u) = .5$.

Figure 5

Participants (bars) plus full *model with fitted* τ_1, ε, κ *and* τ_2 *(black points). Blue highlights for canonical "everything happened" scenario (*$A = 1$, $B = 1$, $E = 1$*). Bars show proportion of responses each variable state received in each world; fill indicates whether the node was observed (orange) or unobserved (green). Error bars show* ± 1 *SE of the observed proportion, colored black for Actual causes and gray for non-Actual causes.*

module is signified by 'no-X' in the name.[8] Model 12 ('everything lesioned') results in observed values of *A* and *B* receiving causal score of 1 and unobserved variables receiving their prior probability; it is thus not identical to the baseline model which assigns uniform probability to all answers.

Figure 5 shows predictions of the full model and participant results for all three probability settings. (Thus Figure 5c shows the same participant results as in Figure 4, here overlaid with the predictions of the full model.) The item-level Pearson correlation coefficient between the full model and participants' average judgments was $r = .92$. The full model had the third best fit, just after two models that had the Actual Causation module lesioned (we discuss the success of the models without Actual Causation later; see 'Summary of results').

Model fit by participant

To calculate a model fit by participant, we split the data by participant and optimized the model fit by minimizing negative log likelihood in the same way as for the aggregate model. We then calculated BIC for each model for each participant and recorded which model had lowest BIC for each participant. We then counted how many participants had each model as their best-fitting model. The results are in the rightmost column of Table 3. The distribution of model fits is not uniform: $\chi^2(12) = 152$, $p < 0.001^{***}$. The individual fits paint a somewhat different picture: The `no Selection` and `no Kindness;no Selection` models fit the largest number of participants, followed by the uniform `Baseline`.

To simplify interpretation, we also looked further at the relative proportions of participants best fit by models with and without each module in turn. We partitioned the set of models four times on whether or not models contained `Kindness`, `Selection`, `Inference`, and `Actual causation` respectively (disregarding those best fit by the uniform `Baseline` model), and then summed the number of participants best fit by the models in each bin. This brings out the relative

[8] Note that if modules were fully independent the power set of the four modules would result in 16 models but since `Kindness` can only act on `Inference`, the four models with kindness but without inference are equivalent to those without both kindness and inference. As a result we fit 12 models.

Table 3

Parameter Values and Performance Metrics for All Models

	Model	τ_1	ε	τ_2	κ	logl	BIC	N participants
1	Full	4.33	.381	.0772	.0925	-4215	8462	13
2	**no ActualCause**	2.44	.228	.0794	.0807	-4174	**8380**	1
3	no Selection		.381	.0709	.0902	-4223	8470	53
4	no ActualCause; no Selection		.236	.0653	.0966	-4246	8515	19
5	no Kindness	2.54	.353	.0932		-4316	8655	9
6	no ActualCause; no Kindness	1.45	.212	.0971		-4276	8577	8
7	no Inference; no Kindness	.375	.378	.6908		-4486	8995	2
8	no Kindness; no Selection		.383	.0743		-4340	8696	30
9	no ActualCause; no Inference; no Kindness	.348	.188	.291		-4673	9370	4
10	no ActualCause; no Kindness; no Selection		.223	.0795		-4520	9056	17
11	no Inference; no Kindness; no Selection		.378	.5857		-4507	9029	8
12	no ActualCause; no Inference; no Kindness; no Selection		.205	.3217		-4956	9928	23
13	Baseline					-5365	10730	28

Note. Model with the lowest Bayesian Information Criterion in bold.

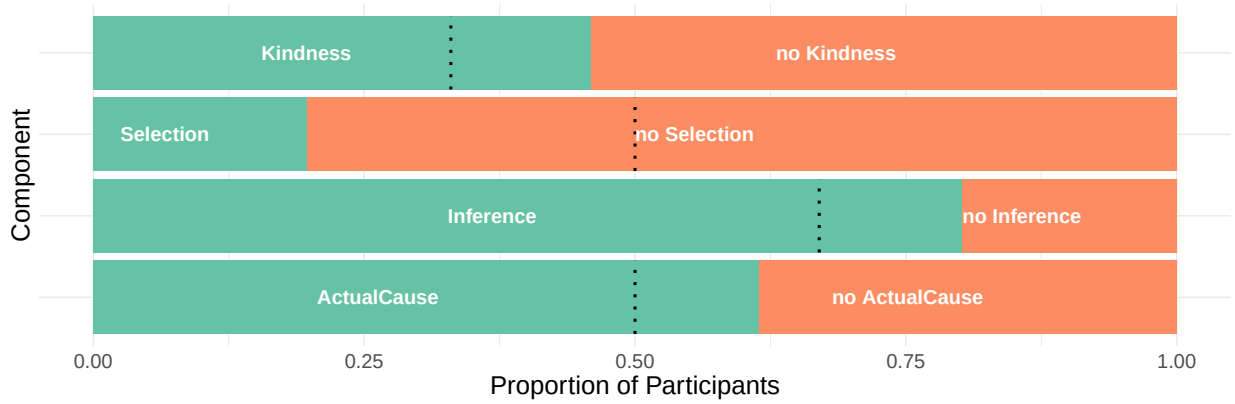


Figure 6

Proportions of participants best fit by models with and without the four modules (excluding the 28 best fit by a uniform baseline). Dotted lines are the proportions expected if participants were allocated to models uniformly.

differences (results shown in Figure 6). We compared the actual numbers of participants to the expected proportions.[9] We ran a simple χ^2 for each module for whether the counts in each bin were different from chance. From bottom up on the chart: More participants were fit by models containing the `Actual causation` module than expected under uniform distribution (61% where random chance would be 50%, $\chi^2(1) = 9.8, p = .002^{**}$). More participants were fit by models containing the `Inference` module than expected under uniform distribution (80% where random chance would be 67.8%, $\chi^2(1) = 15.4$, $p < .001^{***}$). Fewer participants were fit by models containing the `Selection` module than expected under uniform distribution (20% where random chance would be 50%, $\chi(1)^2 = 68.3, < .001^{***}$). More participants were fit by models containing the `Kindness` module than chance allocation (46% where random chance would be 33.3%, $\chi^2(1) = 13.6, p < .001^{***}$).

Alternative measures of Kindness

In Appendix D we conduct a model comparison with alternative, information-theoretic measures of Kindness. We find that none of these measures improves upon the fit of our simple operationalization of Kindness as a constant bonus to unobserved variables (lowest BIC among models with alternative measures: 8450; lowest BIC among models with constant-bonus model of Kindness: 8380).

Summary of results

Our model comparison suggests that inference, selection and kindness modules all helped explain participants' choices in our task *overall*, but that there is considerable heterogeneity in the degree to which the components contribute to individuals choices. Note that because each participant makes only 12 choices, BIC penalization heavily favors simpler models at the individual level. The complexity penalty affects the Kindness and Selection modules in particular

[9] Expected proportions for `Inference` and `Kindness` are more complex than the simple halves of `ActualCause` and `Selection`: eight of the twelve models contained inference and four did not, and that four of the models contained kindness and eight did not. Also note `Kindness` is nested in `Inference` and so those two modules are not strictly dissociable.

(each of which introduces a free parameter), contributing to these modules fitting a relatively small proportion of participants at the individual level.

We now discuss the contribution of each module in more detail. Evidence supporting each module comes from at least three sources: 1) presence of the module in the best-fitting overall model on BIC; 2) comparing aggregate fit for models with and without the module keeping all else equal; 3) number of participants best fit by models with or without the module in question. We summarize and interpret each module in turn on these metrics.

The best-fitting model contained the `Selection` module. Each model containing the `Selection` cognitive module fit better on BIC than the equivalent model without it (for example, `no Kindness` fit better than `no Kindness; no Selection`, etc.). However, far fewer individual participants were fit by models containing the `Selection` cognitive module (37) than without (150)—a result partly explained by the strength of the complexity penalty at the individual level (as discussed above).

The best-fitting model contained the `Inference` cognitive module. Each model containing it fit better on BIC than the equivalent model without it (for example, `no Kindness; no Selection` fit better than `no Inference; no Kindness; no Selection`, etc.). Far more participants were fit by models containing the `Inference` cognitive module (150) than without (37).

The best-fitting model contained the `Kindness` cognitive module, and each model containing it fit better on BIC than the equivalent model without it (for example, `no ActualCause` fit better than `no ActualCause; no Kindness`, etc.). Less than half of participants were fit by models containing the `Kindness` cognitive module (86) than without (101), although this number was still significantly higher than would be expected under random model allocation).

The patterns for the `ActualCause` module are more complex. The best-fitting model *did not* contain the `ActualCause` module. Comparing equivalent models with and without gives a conflicting picture where including the module sometimes does improve fit: for example, `no`

`Inference; no Kindness` fit better than `no ActualCause; no Inference; no Kindness`. Significantly more participants were fit by models containing the `ActualCause` cognitive module (115) than without (72).

The aggregate model fitting suggests actual causality contributed little to causal explanation in this task. The best-fitting model was narrowly the one lesioned to remove the actual causation module: this model gave causal scores even to variables taking values opposite to the effect (e.g., scoring *B*'s absence as a possible explanation of *E*'s presence). These scores were generally negative prior to softmaxing, resulting in a pattern of low but variable selection probabilities across non-actual causes. It is likely some participants ended up selecting non-actual causes non-uniformly because of inattention, even after investing time and resources in computation (that is, they were still influenced by their posterior probabilities and CES scores). This could explain why this erroneous selection of non-actual causes was not entirely modeled by the noise parameter ε, which assigns a small probability to all responses.

Qualitative analysis: behavioral patterns per module

Our model comparison provides evidence that inference, selection and a preference for inferred variables all play a role in how participants choose explanations. Here we illustrate these results, by giving examples of qualitative patterns in participants' choices that can be explained by these three modules.

In order to see whether participants use a particular module in their explanations, a useful strategy is to contrast the predictions of two computational models: one that uses that module, and one that does not. For example, to evaluate whether participants engage in inference, we can compare a model that engages in inference to a model that is identical in all respects to the first model except that it does not use inference. In this section we use this targeted comparison strategy. For better comparability, all model predictions shown in this sub-section use the same values for the free parameters—in contrast these parameters were fit to each model separately in the previous analysis.[10] Note that here we focus on simple effects that arise from the application

[10] Specifically, model predictions displayed in the current sub-section use $\tau_1 = .1$, $\tau_2 = 1$, $\kappa = 1$ and $\varepsilon = 0$.

of one module, but these modules also interact with each other in ways that are more difficult to illustrate.

Do participants engage in inference?

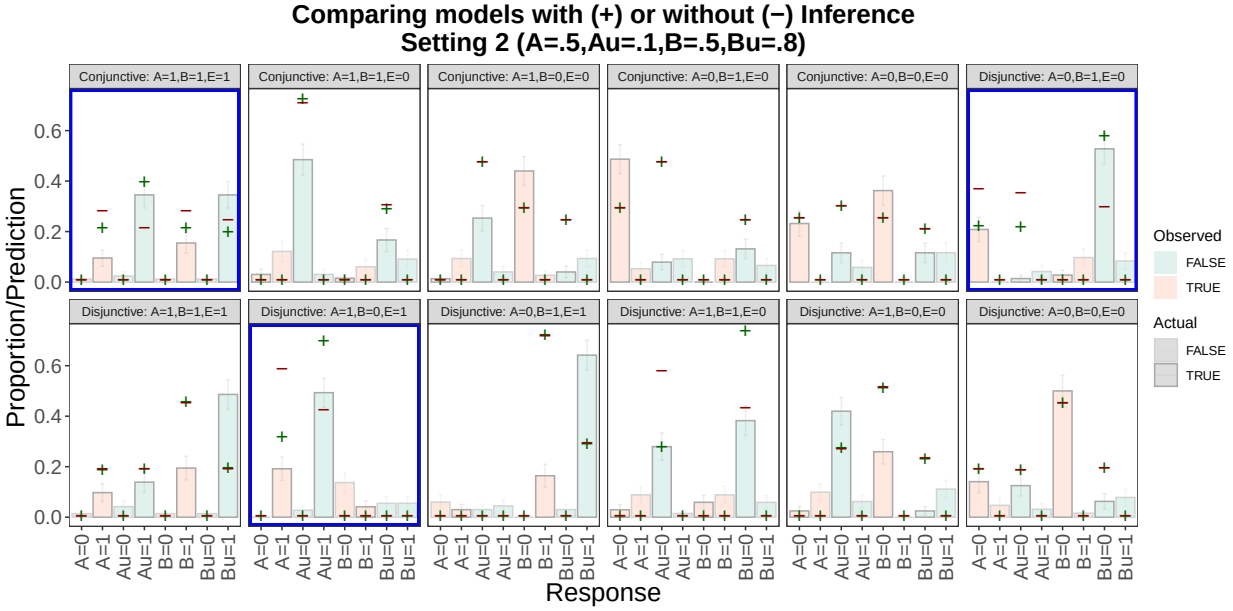


Figure 7

Comparison between predictions of a model with inference (green plus) and without inference (red minus), overlain on participant bars. Both computational models have all other modules except for Kindness. Blue borders highlight trials mentioned in the main text. Bars show proportion of responses each variable state received in each world; fill indicates whether the node was observed (orange) or unobserved (green).

Figure 7 displays the predictions of two versions of our model in probability setting 2: one with and one without the Inference module. Overall the version of the model with Inference attributes higher scores to unobserved variables than the model without Inference does. This tendency holds when the value of the unobserved variables can be inferred from the observed events. For example in the Conjunctive trial with $A = 1, B = 1, E = 1$ (see top left facet of Figure 7), the model without Inference (red minus signs) favors $A = 1$ over $A_u = 1$, but the model with Inference favors $A_u = 1$ (see green plus signs) because the observations license the inference

that A_u must have happened. Participants show the same preference as the Inference model, where 29 chose $A_u = 1$ compared with 8 who chose $A = 1$ ($\chi^2(1) = 11.9$, $p < .001^{***}$). Similarly, in the Disjunctive trial with $A = 1$, $B = 0$ and $E = 1$, the model without Inference favors $A = 1$ over $A_u = 1$, but the model with Inference favors $A_u = 1$. Again, participants show the same preference, where 36 chose $A_u = 1$ compared with 14 who chose $A = 1$ ($\chi^2(1) = 9.7$, $p = .0019^{**}$).

The model with Inference can also sometimes explain why an unobserved variable is preferred to another unobserved variable. In the Disjunctive trial with $A = 0$, $B = 1$, $E = 0$, participants strongly preferred to explain why E did not happen by mentioning $B_u = 0$ than $A_u = 0$, where 38 chose $B_u = 0$ compared with 1 who chose, $A_u = 0$ ($\chi^2(1) = 35.10$, $p < .001^{***}$). This preference cannot be explained by the model without Inference, but it can be explained by the model with Inference. The observation that $A = 0$, $B = 1$, $E = 0$ licenses the inference that B_u must have failed to happen, whereas we cannot make any inference about whether A_u happened. Therefore the model with inference prefers to mention $B_u = 0$, whose state is now certain, relative to $A_u = 0$, whose state is still unknown.

Do participants engage in counterfactual-based selection?

Here we examine three behavioral signatures of counterfactual-based causal selection. To better visualize these signatures, on Figure 8 and Figure 9 we plot the predictions of two versions of our computational model, that are identical to each other except that one engages in counterfactual-based causal selection and the other does not.

Sensitivity to event probability. According to the CES model, the prior probability of an event should have an effect on whether this event is mentioned in explanations, all else being equal (Quillien & Lucas (2023); see also Kahneman & Miller (1986); Icard et al. (2017)). We can examine this prediction in probability setting 1, where $P(A) < P(B)$ but $P(A_u) = P(B_u)$, and in probability setting 2, where $P(A) = P(B)$ but $P(A_u) < P(B_u)$.

In probability setting 1 (see Figure 8), we compare the relative frequency of explanations that mention A vs B. We first look at trials where all observable events happen ($A = 1$, $B = 1$, $E = 1$). In the conjunctive condition participants choose to mention $A = 1$ more than $B = 1$, with

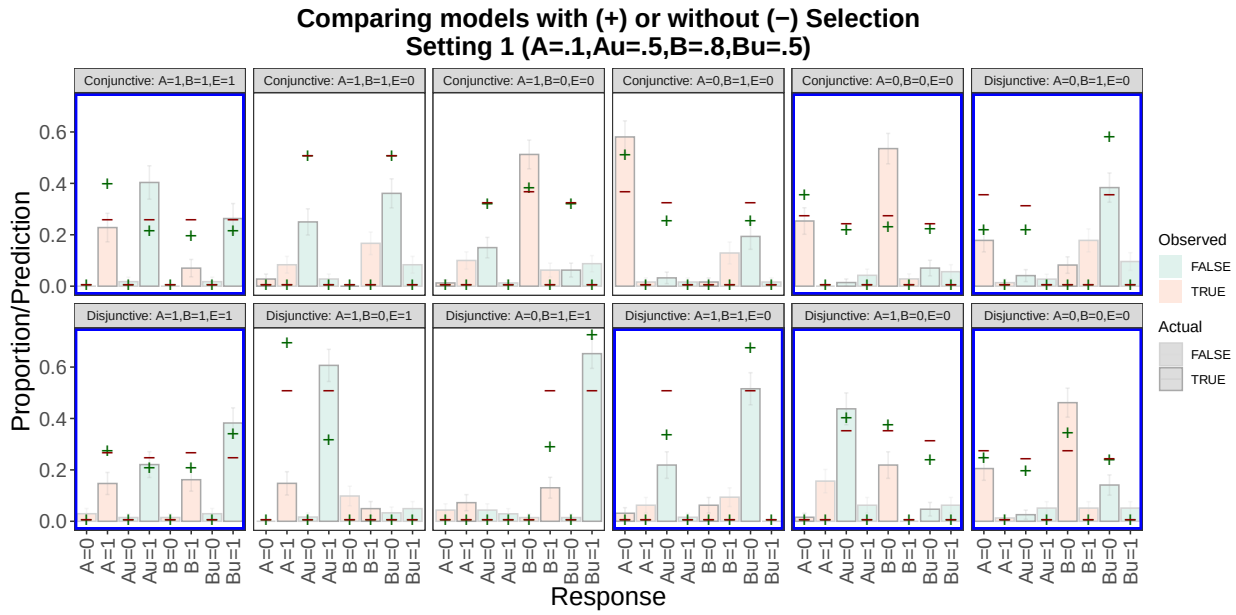


Figure 8

Comparison between predictions of a model with selection (green plus) and without selection (red minus), overlain on participant bars. Both computational models have all other modules except for Kindness. Blue borders highlight trials mentioned in the main text. Bars show proportion of responses each variable state received in each world; fill indicates whether the node was observed (orange) or unobserved (green).

13 participants choosing $A = 1$ compared with 4 choosing $B = 1$ ($\chi^2(1) = 4.76$, $p = .029^*$), but show no such preference in the disjunctive trial, where 10 chose $A = 1$ compared with 11 who chose $B = 1$ ($\chi^2(1) = .048$, $p = .83$). This pattern is partly consistent with the counterfactual model, which predicts a preference for $A = 1$ on both trials but does so more strongly in the conjunctive trial.[11] We then look at trials where none of the observable events happen ($A = 0$, $B = 0$, $E = 0$). Participants select the low-probability event $B = 0$ more often than they select $A = 0$ both in the conjunctive trial (38 compared with 18, $\chi^2(1) = 7.14$, $p = .0075^{**}$) and the disjunctive trial (36 compared with 16, $\chi^2(1) = 7.69$, $p = .0055^{**}$). However, the counterfactual

[11] Note that in simpler disjunctive structures counterfactual models would predict that low-probability events like A should be seen as *less* causal, not more (Icard et al., 2017; Quillien, 2020). The CES model does not make this prediction here because the disjunctive structure is more complicated, due to the presence of the unobserved variables.

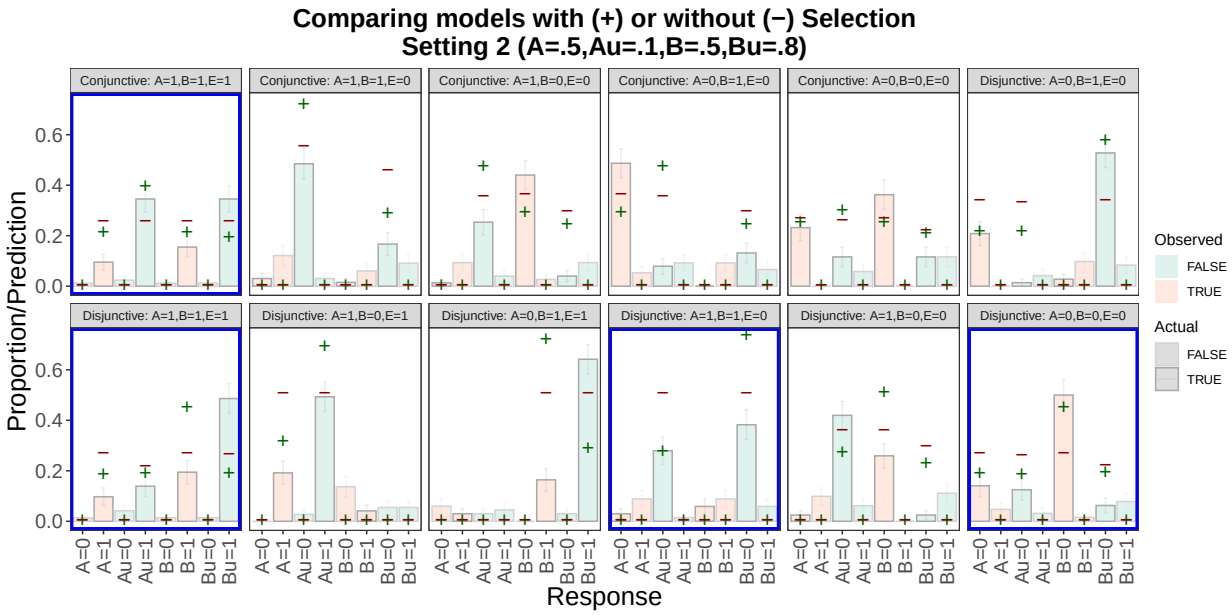


Figure 9

Comparison between predictions of a model with selection (green plus) and without selection (red minus), overlain on participant bars. Both computational models have all other modules except for Kindness. Blue borders highlight trials mentioned in the main text. Bars show proportion of responses each variable state received in each world; fill indicates whether the node was observed (orange) or unobserved (green).

model predicts this effect only for the disjunctive trial, and predicts the inverse effect for the conjunctive trial.

In probability setting 2 (see Figure 9), $P(A_u) < P(B_u)$ while A and B have equal probability, so we can look at whether people mention one of the unobserved events more than the other, all else being equal. The relevant trials are the two trials where the values of A_u and B_u can be inferred with certainty (in these trials uncertainty about the variables' actual-world values is not a confound). In the conjunctive trial with $A = 1$, $B = 1$, $E = 1$, we can infer with certainty that both A_u and B_u happened. The counterfactual model predicts a preference for low-probability event $A_u = 1$ over $B_u = 1$, but participants do not show a preference either way in this trial (both 29, $\chi^2(1) = 0$, $p = 1$). In the disjunctive trial with $A = 1$, $B = 1$, $E = 0$, we can infer with

certainty that neither A_u nor B_u happened. The counterfactual model predicts a preference for low-probability event $B_u = 0$, and participants exhibit the same preference mildly, but the difference is not significant (26 compared with 19, $\chi^2(1) = 1.09$, $p = .297$).

Sensitivity to the probability of other events. According to the CES model, the prior probability of *other* events should have an influence on whether an event is mentioned in an explanation (see also Kominsky et al., 2015; Icard et al., 2017). In our scenarios, the model makes this prediction because an event that can cause an outcome only if a low-probability event also happens will typically not be highly correlated with the outcome across counterfactual possibilities.

In probability setting 1 (see Figure 8), we compare the relative frequency of mention of A_u and B_u: they have the same prior probability in this setting, but B_u is associated with a variable that has higher probability (i.e. $P(B) > P(A)$). Again, we look at the trials where the value of A_u and B_u can be inferred with certainty. In one of these trials, the disjunctive trial with $A = 1$, $B = 1$, $E = 0$, the counterfactual model favors $B_u = 0$ over $A_u = 0$. Participants show the same preference (33 compared with 14, $\chi^2(1) = 7.68$, $p = .0056^{**}$).

In probability setting 2, we compare the relative frequency of mention of A and B: they have the same probability in this setting, but the causal influence of B is more likely to 'work' (i.e. $P(B_u) > P(A_u)$). There are two trials where the counterfactual model predicts a difference between A and B while the inference module assigns the same posterior probability to both events. In the disjunctive trial with $A = 1$, $B = 1$, $E = 1$, the model favors $B = 1$, and participants show the same preference mildly, but the difference is not significant (14 compared with 7, $\chi^2(1) = 2.33$, $p = .127$). In the disjunctive trial with $A = 0$, $B = 0$, $E = 0$, the model favors $B = 0$, and participants show the same preference (where 32 chose $B = 0$ compared with 9 who chose $A = 0$, $\chi^2(1) = 19.90$, $p < .001^{***}$).

Sensitivity to actual necessity. The CES model also holds that counterfactual simulation is 'anchored' to the actual world: people rarely simulate counterfactual possibilities that are too dissimilar to what actually happened (Quillien & Lucas, 2023). Such a bias predicts a specific

empirical signature: People should tend to mention in their explanations events that, in the actual world, were necessary for the outcome to happen (see also Chockler & Halpern, 2004; Lagnado et al., 2013; Icard et al., 2017). An event is necessary for the outcome in this sense if a counterfactual intervention undoing this event would have been enough to prevent the outcome from happening.

We can test this prediction in probability setting 1, where the unobserved variables have equal prior probability. In some of these trials, only one of the observed variables happened, and so the associated unobserved variable would have flipped the outcome if its value had changed. For example, in the disjunctive trial with $A = 0$, $B = 1$, $E = 0$, we can infer that $B_u = 0$, and that E would have happened if B_u had been 1. In contrast, flipping the value of A_u would have made no difference to whether the outcome happened. In this trial, the counterfactual model predicts a preference for $B_u = 0$ over $A_u = 0$, and participants show the same preference (where 28 chose $B_u = 0$ compared with 3 who chose $A_u = 0$, $\chi^2(1) = 20.16$, $p < .001^{***}$). Similarly, in the disjunctive trial with $A = 1$, $B = 0$ and $E = 0$, the counterfactual model has a preference for $A_u = 0$ relative to $B_u = 0$, and participants show the same preference (also, coincidentally, 28 compared with 3, $\chi^2(1) = 20.16$, $p < .001^{***}$). Note, however, that on these trials participants' preference is also predicted (to a lesser extent) by the inference component of the model.

In sum, many patterns in participants' explanation choice are consistent with counterfactual-based selection, although not all predictions of the counterfactual model are successful.

Do participants favor inferred variable states?

We can compare the predictions of our model with and without the kindness module which adds a constant bonus to inferred variable states—see Figure 10. We focus on probability setting 1 because here A_u and B_u have the same prior probability and so any effect of scaling inference will be more comparable than in either of the other two settings. There are two trials where the model without kindness favors observed variables while the model with kindness favors unobserved variables. In the conjunctive trial with $A = 1$, $B = 1$, $E = 1$, the model without

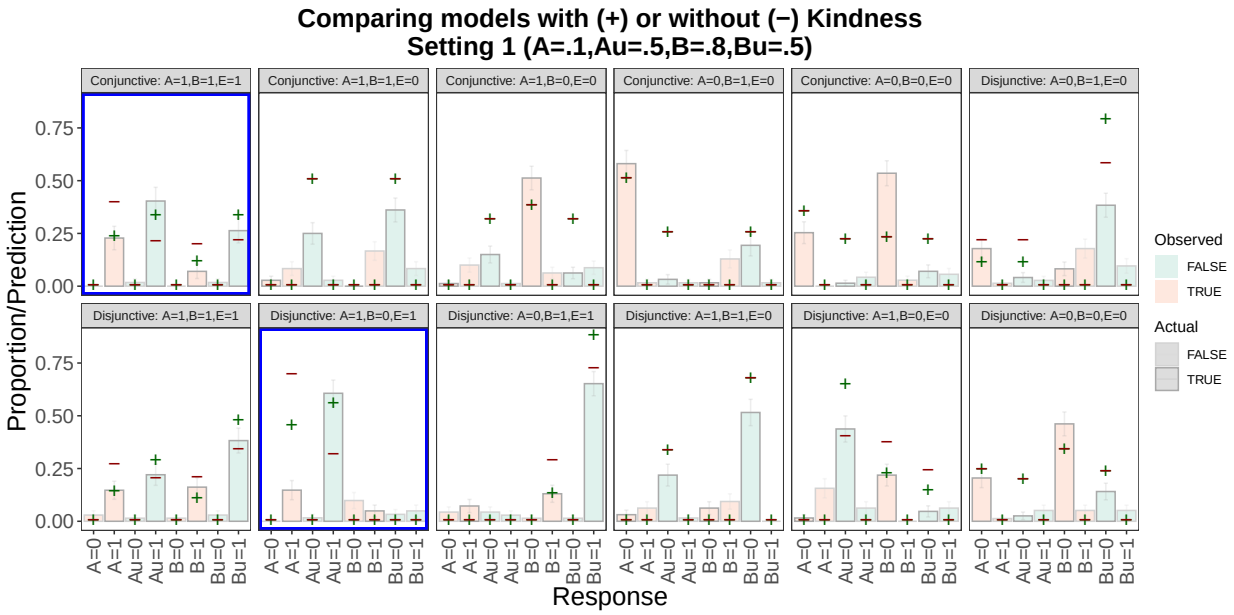


Figure 10

Comparison between predictions of a model with kindness (green plus) and without kindness (red minus), overlain on participant bars. Both computational models have all other modules. Blue borders highlight trials mentioned in the main text. Bars show proportion of responses each variable state received in each world; fill indicates whether the node was observed (orange) or unobserved (green).

kindness favors $A = 1$ over $A_u = 1$ and shows no preference between B and B_u. In contrast, the model with the kindness module prefers the unobserved variables in this trial, because the observed events license the inference that both A_u and B_u must have happened. Participants show the same preference, with 40 choosing an unobserved variable compared with 17 choosing an observed variable ($\chi^2 = 9.28$, df = 1, $p = .0023^{**}$). In the disjunctive trial with $A = 1$, $B = 0$, $E = 1$, the model without kindness favors $A = 1$ over $A_u = 1$. But because the observed events license the inference that A_u must have happened, the model with the kindness module favors $A_u = 1$ over $A = 1$. Participants show the same preference, where 37 chose $A_u = 1$ compared with 9 who chose $A = 1$ ($\chi^2(1) = 17.04$, $p < .001^{***}$).

Figures illustrating the relative role of each module in all probability settings can be found

in Appendix C.

Discussion

Causal explanation is a complex cognitive activity that requires solving multiple sub-problems. Research on causal cognition has typically focused on one sub-problem at a time: some experiments focus on causal inference while other experiments focus on causal selection. Although this strategy has been fruitful so far, if inference and selection interact in real-world causal explanation then studying them in isolation can only lead to a limited understanding. Here we study these two problems together. First, we sketched a computational framework for how these two processes can be integrated. We expanded this computational framework with the concept of computational kindness, to lay groundwork for an account of how inference and selection interact in situations of uncertainty and potential for communicative demand. Then we reported the results of an experiment testing how people give causal explanations under uncertainty. Our computational model comparison supported the idea that human explanations recruited causal selection and causal inference. Our exploratory analyses suggested that participants also favored mentions of unobserved (i.e., inferred) components more than we would expect if inference and selection were all that were in play.

Our experiment suggests people engage in causal selection in a way that is consistent with existing work. We implemented a computational model of causal selection, the Counterfactual Effect Size Model (Quillien, 2020; Quillien & Lucas, 2023). Under this model, people perform causal selection by simulating other possible outcomes similar to what actually happened (‘counterfactuals’), and then assessing how robustly the candidate cause correlates with the observed effect across the counterfactuals. We found support for causal selection like that modeled by Quillien & Lucas (2023) through its inclusion in our best fitting model. Our results provide a timely extension to the existing literature on this model (Quillien & Barlev, 2022; O’Neill et al., 2025; Konuk et al., 2026) by showing it can account for people’s explanatory behavior in richer settings containing some uncertainty through unobserved variables.

Our modeling also suggested that people were engaging in inference, assessing the

evidence-conditioned probabilities of variables' states and marginalizing over these probabilities when making their selection. Our results enrich work on inferences from explanations by Kirfel et al. (2022) and others by capturing a complementary way in which inference and explanation interact even without any overt information asymmetry between an explainer and receiver. In addition, our formalization efforts contribute to the unification of inference and judgment found in causal cognition research, and so supports causal model theory as a productive framework within which to study explanatory behavior.

Many participants in our experiment tended to select inferred variable states over and above what one would expect purely based on Bayesian inference and selection mechanisms. We proposed that such a preference might be grounded in a general motivation to favor explanations that package the products of the work the explainer has done. We operationalized this in our primary analysis as a simple constant bonus for unobserved variables. However we also explored more complex, information-theoretic measures as potential alternatives (see Appendix D). This additional model comparison revealed that none of these measures outperformed the simple inferred-variable bonus component. Our experiment was not set up to test this aspect of participants' choices, and the addition of the descriptive kindness component was post-hoc, so we do not draw any firm conclusions from these particular results. The communicative value of explanations under uncertainty is an important avenue for follow-up work (cf. Harding et al., 2025). For example, while our explanation task did not have an explicitly social aspect, future research could introduce an audience, and manipulate what information the audience has access to (Kirfel et al., 2024). It would be valuable to combine this with direct manipulations of the cognitive cost of inference, allowing for development of more fine-grained theories of the way that explainers package the products of their cognitive efforts to benefit a listener.

Conclusion

In this paper we studied how people select causal explanations about singular events in a setting where the value of some variables is unobserved. We developed a computational model that incorporated modules for causal inference and causal selection, and found that both of these

modules, as well as preference for inferred variables, helped to account for people's aggregate explanatory preferences. Our work contributes a fuller picture of how people decide what makes for a good explanation in an uncertain and incompletely observed world.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Baumgartner, M., & Glynn, L. (2013). Introduction to special issue on 'actual causation'. *Erkenntnis*, *78*(Suppl 1), 1–8.

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing neurath's ship: Approximate algorithms for online causal learning. *Psychological review*, *124*(3), 301.

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 708.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological review*, *104*(2), 367.

Chockler, H., & Halpern, J. Y. (2004). Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, *22*, 93–115.

Christian, B., & Griffiths, T. (2016). *Algorithms to live by: The computer science of human decisions*. Macmillan.

Coenen, A., Rehder, B., & Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive psychology*, *79*, 102–133.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory* (Vol. 2). wiley New York.

Davis, Z. J., & Rehder, B. (2020). A process model of causal reasoning. *Cognitive Science*, *44*(5), e12839.

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*, 1–12.

Dijkstra, E. W. (2002). Ewd1300: The notational conventions i adopted, and why. *Formal aspects of Computing*, *14*(2), 99–107.

Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *Journal of experimental psychology: Learning, memory, and cognition*, *35*(3), 678.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*.

Gerstenberg, T., & Icard, T. (2020). Expectations affect physical causation judgments. *Journal of Experimental Psychology: General*, *149*(3), 599.

Gerstenberg, T., Ullman, T. D., Nagel, J., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2018). Lucky or clever? from expectations to responsibility judgments. *Cognition*, *177*, 122–141.

Gopnik, A., Schulz, L., & Schulz, L. E. (2007). *Causal learning: Psychology, philosophy, and computation*. Oxford University Press.

Hagmayer, Y., Sloman, S. A., Lagnado, D. A., & Waldmann, M. R. (2007). Causal reasoning through intervention. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 86–100). Oxford University Press.

Halpern, J. Y. (2016). *Actual causality*. MiT Press.

Halpern, J. Y., & Pearl, J. (2005a). Causes and explanations: A structural-model approach. Part I: Causes. *The British Journal for the Philosophy of Science*.

Halpern, J. Y., & Pearl, J. (2005b). Causes and explanations: A structural-model approach. Part II: Explanations. *The British Journal for the Philosophy of Science*.

Harding, J., Gerstenberg, T., & Icard, T. (2025). A communication-first account of explanation. *arXiv preprint arXiv:2505.03732*.

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.

Hesslow, G. (1988). The problem of causal selection. In D. J. Hilton (Ed.), *Contemporary science and natural explanation: Commonsense conceptions of causality.* New York University Press. doi: 10.1086/355318

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, *107*(1), 65.

Hitchcock, C. (2007). Prevention, preemption, and the principle of sufficient reason. *The Philosophical Review*, *116*(4), 495–532.

Hitchcock, C., & Knobe, J. (2009). Cause and norm. *The Journal of Philosophy*, *106*(11), 587–612.

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual review of psychology*, *62*, 135–163.

Ibeling, D., & Icard, T. (2023). Comparing causal frameworks: Potential outcomes, structural models, graphs, and abstractions. *Advances in Neural Information Processing Systems*, *36*, 80130–80141.

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80–93.

Janzing, D., & Schölkopf, B. (2010). Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, *56*(10), 5168–5194.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological review*, *93*(2), 136.

Kinney, D., & Lombrozo, T. (2024). Building compressed causal models of the world. *Cognitive Psychology*, *155*, 101682.

Kirfel, L., Harding, J., Shin, J. Y., Xin, C., Icard, T., & Gerstenberg, T. (2024). Do as i explain: Explanations communicate optimal interventions. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

Kirfel, L., Icard, T., & Gerstenberg, T. (2022). Inference from explanation. *Journal of Experimental Psychology: General*, *151*(7), 1481.

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2015). Causal superseding. *Cognition*, *137*, 196–209.

Konuk, C., Quillien, T., & Mascarenhas, S. (2026). Plural causes. *Open Mind*, *10*, 441–477.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, *22*(1), 79–86.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. doi: 10.18637/jss.v082.i13

Lagnado, D. A. (2021). *Explaining the evidence: How the mind investigates the world*. Cambridge University Press.

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science*, *37*(6), 1036–1073.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*(10), 464–470.

Lombrozo, T., & Vasilyeva, N. (2017). Causal explanation. *The Oxford handbook of causal reasoning*, 415.

Lucas, C. G., & Kemp, C. (2015). An improved probabilistic account of counterfactual reasoning. *Psychological Review*, *122*(4), 700.

Meder, B., & Mayrhofer, R. (2017). Diagnostic causal reasoning with verbal information. *Cognitive Psychology*, *96*, 54–84.

Meteyard, L., & Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*, 104092.

Morris, A., Phillips, J., Gerstenberg, T., & Cushman, F. (2019). Quantitative causal selection patterns in token causation. *PloS one*, *14*(8), e0219704.

Nam, A., Hughes, C., Icard, T., & Gerstenberg, T. (2023). Show and tell: Learning causal structures from observations and explanations. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

Navarre, N., Konuk, C., Bramley, N. R., & Mascarenhas, S. (2024). Functional rule inference from causal selection explanations. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

Nelson, J. D. (2005). Finding useful questions: on bayesian diagnosticity, probability, impact, and information gain. *Psychological review*, *112*(4), 979.

Nielsen, U., Pellet, J.-P., & Elisseeff, A. (2012). Explanation trees for causal bayesian networks. *arXiv preprint arXiv:1206.3276*.

O'Neill, K., Henne, P., Quillien, T., Icard, T., & DeBrigard, F. (2025). Norms moderate causal judgments in cases of double prevention. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 47).

Pacer, M., Williams, J., Chen, X., Lombrozo, T., & Griffiths, T. (2013). Evaluating computational models of explanation using human judgments. In *Proceedings of the twenty-ninth conference on uncertainty in artificial intelligence.*

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge: Cambridge University Press.

Phillips, J., Luguri, J. B., & Knobe, J. (2015). Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition*, *145*, 30–42.

Quillien, T. (2020). When do we think that x caused y? *Cognition*, *205*, 104410.

Quillien, T., & Barlev, M. (2022). Causal judgment in the wild: evidence from the 2020 us presidential election. *Cognitive Science*, *46*(2), e13101.

Quillien, T., Bramley, N., & Lucas, C. G. (2025). Lossy encoding of distributions in judgment under uncertainty. *Cognitive Psychology*, *159*, 101745.

Quillien, T., & Lucas, C. G. (2023). Counterfactuals and the logic of causal selection. *Psychological Review*.

Quillien, T., Szollosi, A., Bramley, N. R., & Lucas, C. (2023). Causal inference shapes counterfactual plausibility. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

Sloman, S. A. (2022). Causal Bayes nets as psychological theory. In H. Geffner, R. Dechter, & J. Y. Halpern (Eds.), *Probabilistic and causal inference: The works of judea pearl* (1st ed., pp. 853–866). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3501714.3501758

Sloman, S. A., & Lagnado, D. A. (2004). Causal invariance in reasoning and learning. *Psychology of Learning and Motivation*, *44*, 287–326.

Von Kügelgen, J., Mohamed, A., & Beckers, S. (2023). Backtracking counterfactuals. In *Conference on causal learning and reasoning* (pp. 177–196).

Woodward, J. (2021). *Causation with a human face: Normative theory and descriptive psychology*. Oxford University Press.

Ying, L., Hillel, A., Truong, R., Mansinghka, V. K., Tenenbaum, J. B., & Zhi-Xuan, T. (2025). Belief attribution as mental explanation: The role of accuracy, informativity, and causality. *arXiv preprint arXiv:2505.19376*.

Yuan, C., Lim, H., & Lu, T. (2011). Most relevant explanation in bayesian networks. *Journal of Artificial Intelligence Research*, *42*, 309–352.

Zhu, J.-Q., & Griffiths, T. L. (2026). Computation-limited bayesian updating: A resource-rational analysis of approximate bayesian inference. *Psychological Review*, *133*(3), 619.

Appendix A

Collider configurations

Table A1

All Collider Worlds with their Unobserved Variable Settings

Conjunctive

	c1				c2				c3				c4			c5
A	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
A_u	0	1	0	1	0	1	0	1	0	0	1	1	0	1	0	1
B	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
B_u	0	0	1	1	0	0	1	1	0	1	0	1	0	0	1	1
E	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Disjunctive

	d1				d2		d3		d4		d5		d6		d7	
A	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
A_u	0	1	0	1	0	1	0	1	0	0	1	1	0	1	0	1
B	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
B_u	0	0	1	1	0	0	1	1	0	1	0	1	0	0	1	1
E	0	0	0	0	0	0	1	1	0	0	1	1	0	1	1	1

Note. Gray banding of rows is simply to assist visual clarity. *Width* of each gray column banding indicates groups that match in observed variables but vary in unobserved variables: for example, in d6 and c5 both unobserved variables have only one possible value and in d2:d5 the value of one unobserved variable can be deduced but the other could take either value.

Appendix B

Cover stories

The behavioral experiment was presented through three cover stories, shown below. In the experiment the text was spread across several click-through screens with accompanying figures. Square brackets indicate phrases that depended on the condition. Here the cover stories are shown with probabilities from probability setting 3 (A=.1,Au=.7,B=.8,Bu=.5).

Cover story 1: cafe

The situation is ... a person walking down a street full of cafes, looking for a place to eat. They have already decided what they want to eat: a main dish, a dessert, or both. They read the menus, and decide whether to enter and eat something. If the food is on the menu and they want to eat it, they will enter the cafe. [10%] of the cafes have main dishes and [80%] have desserts. The person wants to eat a main dish [70%] of the time and wants to eat a dessert [50%] of the time. The person enters the cafe to order if [either one / both] of the main dish or dessert is on the menu and the person wants to eat what is on the menu.

Cover story 2: class

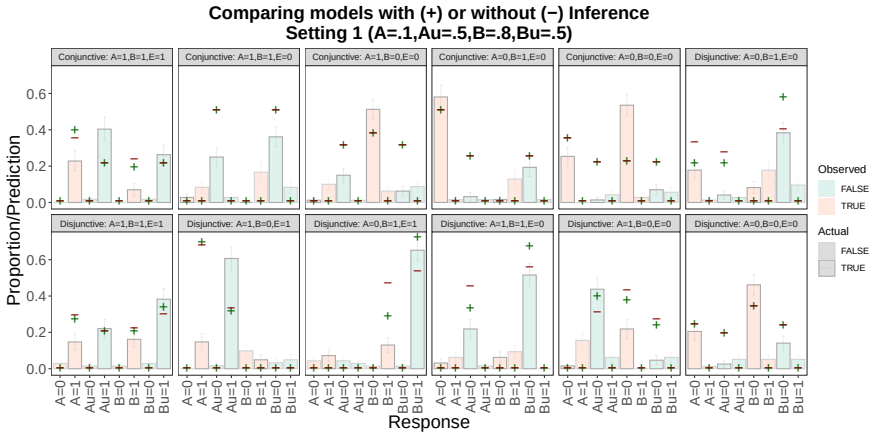
The situation is ... an afternoon university seminar class. The class is compulsory, but two particular students only sometimes attend. These two students are intelligent, passionate, articulate and well-read. However, even when they attend, they are only in the mood to talk if they have had a good morning! (Their morning and hence their mood does not affect how likely they are to turn up; a bad morning just makes them quiet.) But if they attend and had a good morning, there will always be a good discussion. The first student attends [10%] of the time and the second student attends [80%] of the time. The first student has a good morning [70%] of the time, and the second student has a good morning [50%] of the time. A good discussion always happens when [either/both] student attends and had a good morning.

Cover story 3: client

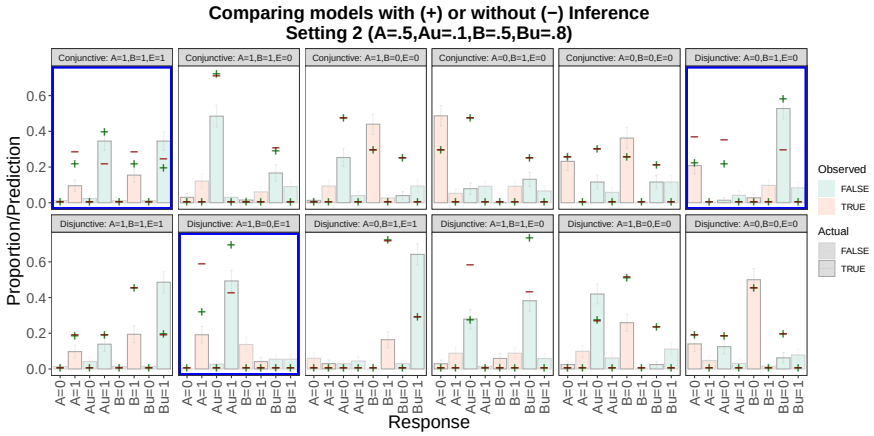
The situation is ... a work meeting, where a company is trying to sell a product to a client. The presenter introduces some features in a talk, and then opens files to demonstrate. However, the laptop files are sometimes randomly corrupted and fail to open! The presenter does not know in advance which files are corrupted. The product has Feature A and Feature B, and the client will always accept the product if they see a working demonstration of [either/both] feature[/s]. The company presents Feature A [10%] of the time and Feature B [80%] of the time. The underlying laptop files for Feature A are sometimes corrupted: they only work [70%] of the time, whether the feature is presented or not. The underlying laptop files for Feature B are sometimes corrupted: they only work [50%] of the time, whether the feature is presented or not.

Appendix C

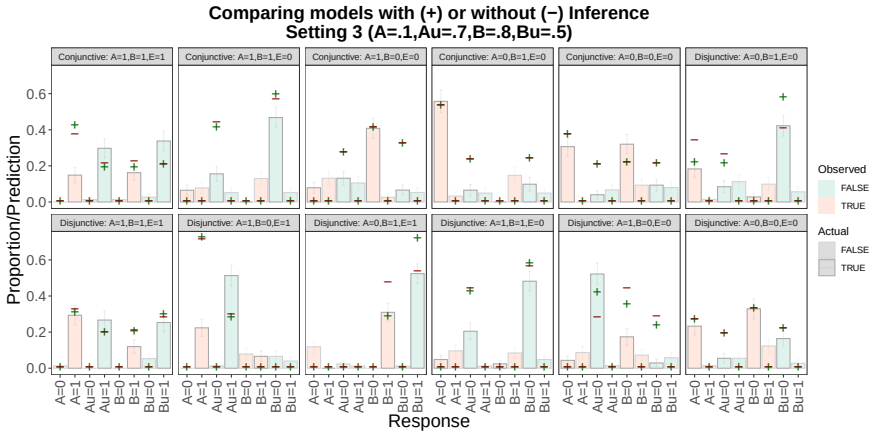
Plots for qualitative behavioral analysis



(a) *Setting 1* $P(A) = .1, P(A_u) = .5, P(B) = .8, P(B_u) = .5$.



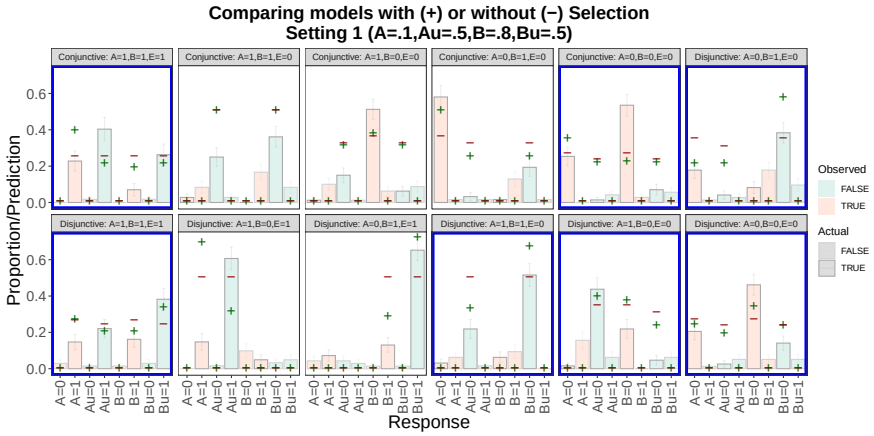
(b) *Setting 2* $P(A) = .5, P(A_u) = .1, P(B) = .5, P(B_u) = .8$.



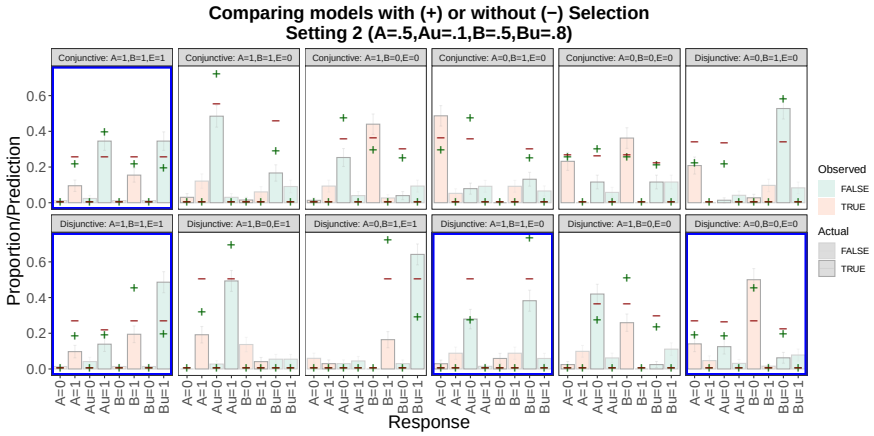
(c) *Setting 3* $P(A) = .1, P(A_u) = .7, P(B) = .8, P(B_u) = .5$.

Figure C1

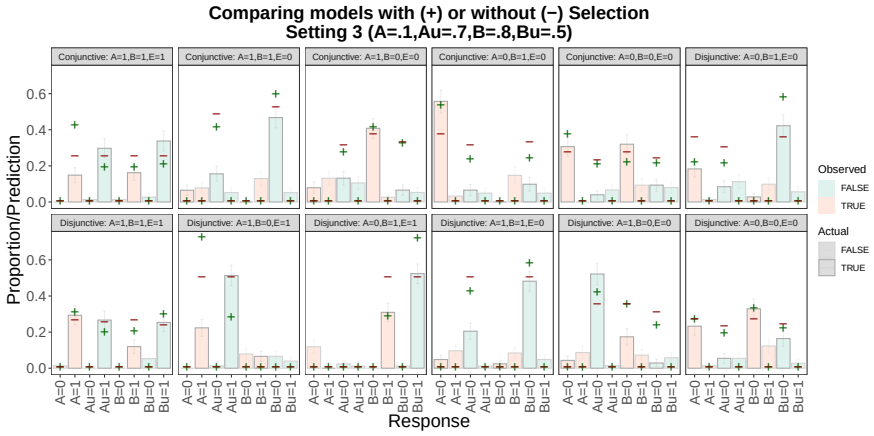
Comparison between predictions of a model with inference (green plus) and without inference (red minus), overlain on participant bars. Both computational models have all other modules except for Kindness. Blue borders highlight trials mentioned in the main text. Bars show proportion of responses each variable state received in each world; fill indicates whether the node was observed (orange) or unobserved (green).



(a) *Setting 1* $P(A) = .1, P(A_u) = .5, P(B) = .8, P(B_u) = .5$.



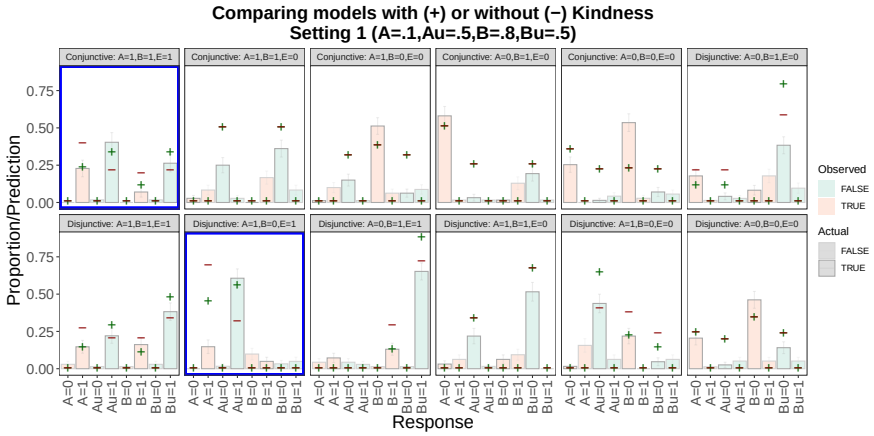
(b) *Setting 2* $P(A) = .5, P(A_u) = .1, P(B) = .5, P(B_u) = .8$.



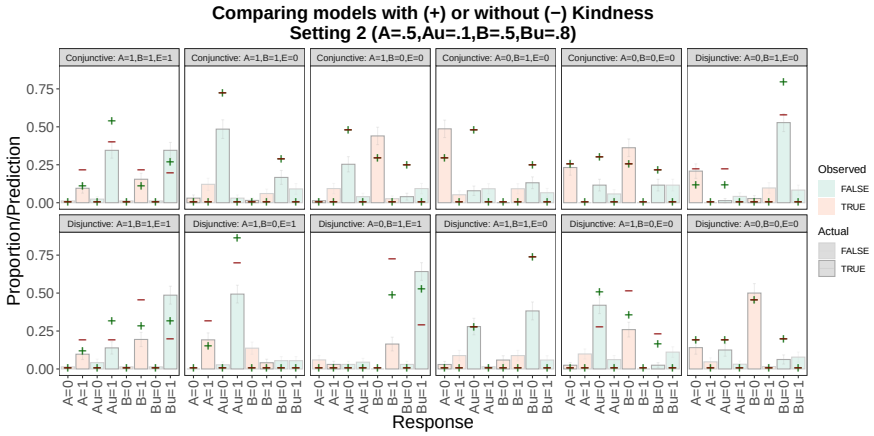
(c) *Setting 3* $P(A) = .1, P(A_u) = .7, P(B) = .8, P(B_u) = .5$.

Figure C2

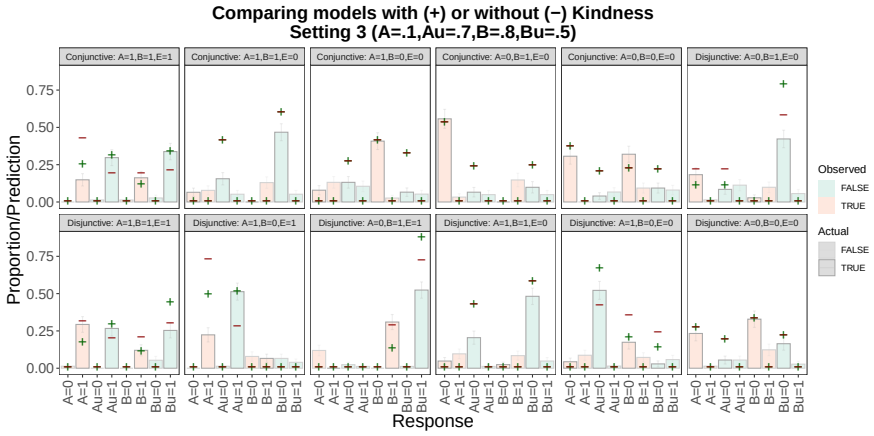
Comparison between predictions of a model with selection (green plus) and without selection (red minus), overlain on participant bars. Both computational models have all other modules except for kindness. Blue borders highlight trials mentioned in the main text. Bars show proportion of responses each variable state received in each world; fill indicates whether the node was observed (orange) or unobserved (green).



(a) *Setting 1* $P(A) = .1, P(A_u) = .5, P(B) = .8, P(B_u) = .5$.



(b) *Setting 2* $P(A) = .5, P(A_u) = .1, P(B) = .5, P(B_u) = .8$.



(c) *Setting 3* $P(A) = .1, P(A_u) = .7, P(B) = .8, P(B_u) = .5$.

Figure C3

Comparison between predictions of a model with Kindness (green plus) and without Kindness (red minus), overlain on participant bars. Both computational models have all other modules. Blue borders highlight trials mentioned in the main text. Bars show proportion of responses each variable state received in each world; fill indicates whether the node was observed (orange) or unobserved (green).

Appendix D

Comparing measures of informative explanation

In our primary analyses we included a constant preference for inferred variables as a constant parameter, finding this improved overall model fit. We were curious as to what specific considerations this preference might originate from for participants. Therefore in this exploratory section we examine several possible measures of computational kindness that, to differing degrees, privilege the inferred over observed variables.

Kullback-Leibler Divergence

Since an explainer must perform inference to obtain posterior probabilities of unobserved variables, we consider one possible operationalization of informativeness as the degree of change between the prior and posterior probability assigned to a mentioned variable. This can be measured using Kullback-Leibler divergence ('KL') (Kullback & Leibler, 1951; Cover & Thomas, 1991).

The relevant KL divergence here is that of posterior to prior since this captures the change with the benefit of hindsight:

$$\mathrm{KL}(P_\alpha(X_u)||P(X_u)) = \sum_{x_u \in X_u} P_\alpha(X_u = x_u) \log_2 \frac{P_\alpha(X_u = x_u)}{P(X_u = x_u)} \tag{D1}$$

This quantity has been used in past work as a measure of information value (Nelson, 2005), and as a measure of the computational cost of belief updating (Zhu & Griffiths, 2026). This KL divergence is positive for any situation in which the observation changes the belief, growing larger the more substantial the change and is zero for the observed variables. We visualize the relative KL scores for all unobserved variables in all conditions in Figure D1.

Information gain

One other salient kindness metric is the change in the level of uncertainty in a variable as a result of inference, often called *information gain*. Intuitively, variables whose state is strongly resolved by inference are potentially good candidates for explanations. Note that information gain is not uniformly positive: variables can occasionally become more uncertain as a result of

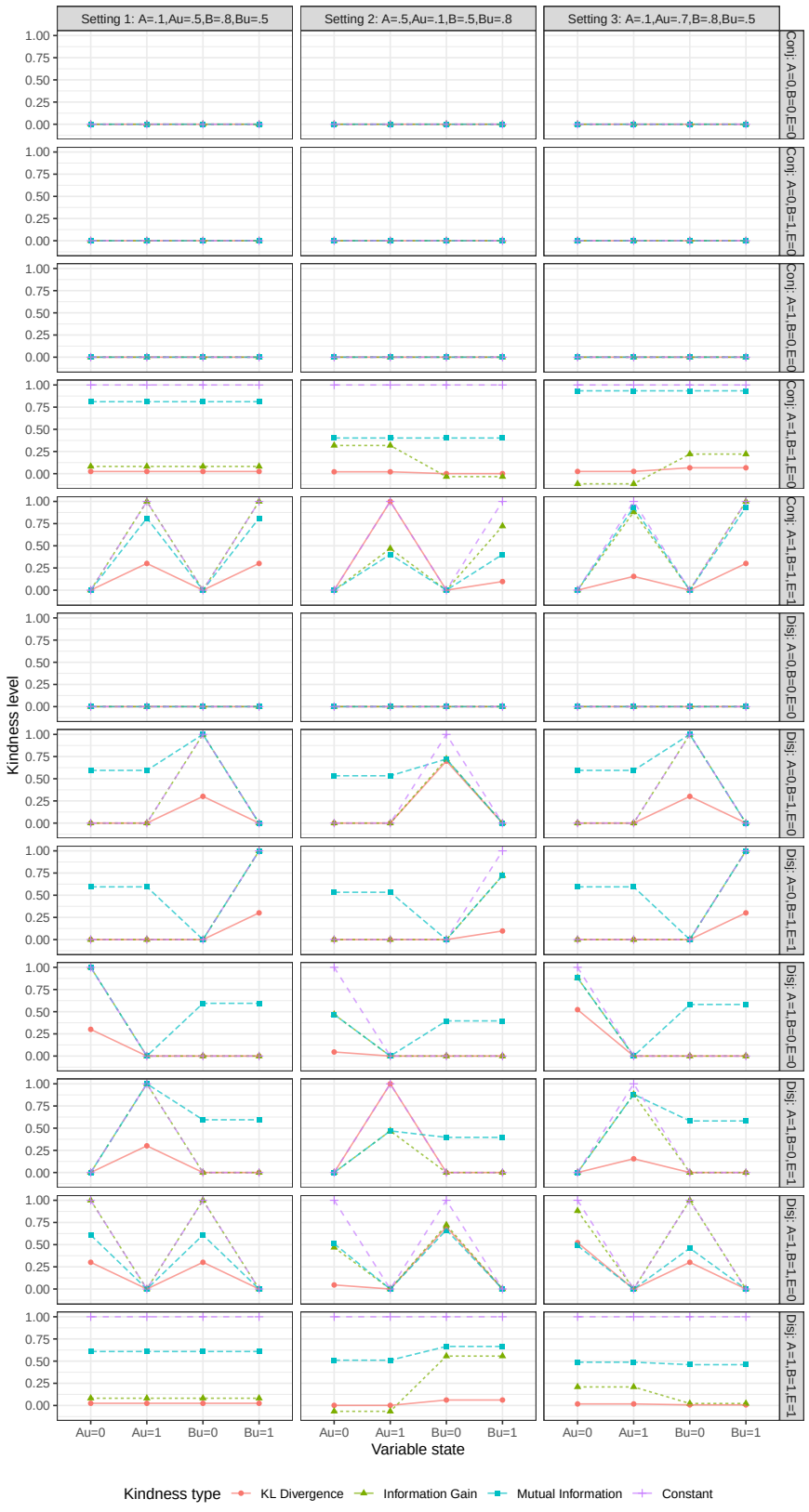


Figure D1

A comparison of the "kindness" bonuses under Kullback-Leibler Divergence (red), Mutual Information (green) and Information gain (Blue). To facilitate comparison all measures are divided by their max score.

inference from the observation of A, B and E (see in Figure D1), complicating its interpretation as a choice bonus, nevertheless we explore whether it gives purchase on participants choices none the less.

Information gain is defined as the reduction in our uncertainty about X_u upon observing the state of the observed variables:

$$\mathrm{IG}(X_u|A=a,B=b,E=e) = H(X_u) - H(X_u|A=a,B=b,E=e), \tag{D2}$$

where $H(\cdot)$ denotes the Shannon entropy:

$$H(X) = -\sum_{x\in X} P(x)\log_2 P(x). \tag{D3}$$

Mutual Information

On the suggestion of an anonymous reviewer we also consider the possibility that preference for observed variables depended on the mutual information between an unobserved variable and the outcome variable, conditioned on the state of the observed variables, i.e. $I(X_u,E|A=a,B=b)$.

Mutual information is computed as:

$$\mathrm{I}(X_u;E \mid A=a,B=b) = \mathbb{E}_{x_u\sim P(X_u|a,b)}\left[\mathrm{IG}(E|x_u)\right] \tag{D4}$$

where

$$\mathrm{IG}(E|x_u) = H\big(E \mid a,b\big) - H\big(E \mid X_u=x_u,a,b\big) \tag{D5}$$

and $H(\cdot)$ denotes the Shannon entropy:

$$H(X) = -\sum_{x\in X} P(x)\log_2 P(x). \tag{D6}$$

Model comparison

We thus define the KL, IG and MI divergences as K_α for each variable and enter these in Equation 5 in place of the constant.

As can be seen in Figure D1, the measures make a variety of different predictions, so are only moderately correlated, with the following Pearson's *r*s: KL-MI= .26; KL-IG= .54;

MI-IG= .62. Table D1 compares the fits of models with each of these exploratory divergence measures against the constant preference for inferred variables (top section). Overall, none of these measures outperformed the constant preference although each still outperformed a no-Kindness model.

Table D1

Parameter Values and Performance Metrics Assuming Different "Kindness" Metrics

	Model	τ_1	ε	τ_2	κ	logl	BIC
	Constant Inferred-variable Bonus						
1	Full	4.33	.381	.0772	.0925	-4215	8462
2	no Act	2.44	.228	.0794	.0807	-4174	**8380**
3	no Select		.381	.0709	.0902	-4223	8470
4	noAct; noSelect		.236	.0653	.0966	-4246	8515
	Kullback-Leibler Divergence						
1	Full	5.5	.381	.0833	.0368	-4286	8603
2	no ActualCause	1.63	.213	.0969	.0232	-4264	8560
3	no Selection		.383	.0774	.0407	-4290	8604
4	no ActualCause; no Selection		.223	.0836	.0564	-4402	8828
	Mutual Information						
1	Full	5.31	.386	.0604	.058	-4269	8569
2	no ActualCause	.00134	.292	2.3759	1.7505	-4725	9482
3	no Selection		.388	.0536	.0575	-4279	8581
4	noActualCause; no Selection		.223	2.2353	1.0808	-4803	9630
	Information Gain						
1	Full	8791.367	.383	.0791	.113	-4247	8526
2	no ActualCause	1.829	.221	.0976	.106	-4209	8450
3	no Selection		.383	.0792	.114	-4248	8519
4	no ActualCause; no Selection		.225	.0868	.15	-4301	8626
5	no Kindness	2.54	.353	.0932		-4316	8655

Note. Columns as in Table 3. See Table 3 for models 6-13.