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).

Data and code are available at our OSF repository, doi.org/10.17605/OSF.IO/TSGJZ.

Inference and selection in causal explanation

Explaining why an event occurred involves solving different information-processing problems: inferring what actually happened (causal inference) and also highlighting which of the causes contributed most significantly to the outcome (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 explanation settings. We also use an information-theoretic measure — information gain — to tease out the relative contribution of the products of inference. We study explanations in scenarios where the states of some variables are unobserved and may be partially or completely inferrable from the outcome, and we also vary the variables' rarity or rates of occurrence. Using a computational model, we show that participants engage in both inference and selection, and also favor explanations that contribute new information. Specifically, participants inferred the state of unobserved variables on the basis of available evidence, selected causes that covary with the outcome across alternative possibilities, and highlighted events whose posterior probability is much higher than their prior probability. Overall, we capture the qualitative patterns in participant selections across a wide range of explanation scenarios with a suite of model variants, suggesting people balance a variety of considerations when providing explanations under uncertainty.

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 our *causal explanations* has received a large amount of attention (Lombrozo, 2006; Lombrozo & Vasilyeva, 2017; Woodward, 2021; Lagnado, 2021; Icard et al., 2017; Quillien, 2020; Quillien & Barlev, 2022; Quillien & Lucas, 2023; Gerstenberg et al., 2021; Hitchcock & Knobe, 2009; Phillips et al., 2015).

Providing a singular causal explanation typically involves solving several different information-processing problems. In this paper we focus on two of the most important:

- **Causal inference.** Causal inference is the process of figuring out what happened on the basis of the available evidence. For example, given the driver was coming back from a party, how likely is it he was drunk? Given it was winter in Canada, how likely is it there was ice on the road?

- **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 know the driver was drunk, that there was ice on the road, and that both factors contributed to the accident. Which will we spotlight as *the* cause of this accident?

It is easy to see that 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 literature, these problems have almost exclusively been studied separately (with a few exceptions introduced 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 make inferences on the basis of observations, interventions and knowledge of causal relations. These inferences may be general or specific, type or token, predictive or diagnostic, hypothetical or counterfactual. Causal inference is often studied using the formalism of causal graphical models (Pearl, 2009). 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), although with noteworthy deviations indicative of process level considerations (Davis & Rehder, 2020).

Causal inference can be at two levels, *type* and *token* inference. Type inference (also called causal discovery or structure learning) covers inference over causal models (cf. Bramley et al., 2015; Coenen et al., 2015). In contrast, *token* inference is a narrower problem of inference over the current value of a variable in a generally-known structure (e.g. Davis & Rehder, 2020). Type inference is a relatively large-scale and stable way to reduce uncertainty, and token inference is more temporally restricted and is often concerned with *events*. Studying token inference involves exploring how people make inferences about whether an event happened, on the basis of information about other events that happened.

A classic token inferential task is to diagnostically infer the state of a potential cause, after observing the effect as well as other potential causes occurring. For example, suppose event C often causes event E, we observe that E happens, and we want to infer the probability that C happened. We can solve this problem using Bayes' rule:

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

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

Actual causation

Even after people have inferred what happened in a particular situation, they still make further judgments about whether one event was the cause of another. These judgments are often referred as judgments of *actual causation*: they are a categorical kind of causal judgment, differentiating between variables that had at least some contribution to an outcome and those that did not contribute at all (Halpern, 2016). It becomes important to include actual causation when modeling relationships that hold across different states of variables. For example, there may be ice on the road, but if the drunk driver did not actually drive on it, then the ice was not an actual cause of the accident. Although actual causation is famously hard to formalize (Baumgartner & Glynn, 2013), technical accounts do exist (Halpern & Pearl, 2005; Halpern, 2016). Since actual causation is not the core topic of our paper, here we will simply use a heuristic definition of actual causation that is adequate for our modeling approach (see Computational framework Section for details).

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 illegal to drive drunk, making the drunk driving an intuitively better explanation as *the* cause of the accident.

For simplicity, extant research on causal selection has usually focused on experimental settings where the reasoner already knows what happened. Because of this, less is known about causal selection in contexts where people also need to make inferences about what happened.

According to a popular family of accounts, people engage in causal selection by imagining other ways events could have turned out: counterfactual possibilities (Icard et al., 2017; Quillien, 2020; Henne et al., 2019), 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 *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 ii) computing a measure of the dependence between *C* and *E* across these possibilities.

The CES model successfully explains data from past experiments on causal judgments (Lagnado et al., 2013; Gerstenberg & Icard, 2020; Icard et al., 2017; Morris et al., 2019). The CES model also made successful new predictions, both in simple experimental settings (Quillien & Lucas, 2023; Konuk et al., 2023; O'Neill et al., 2025) and in a real-world context (Quillien & Barlev, 2022). Although the model fits people's judgments well in fully observable settings, it has not been tested in richer and more complex settings where the state of some variables is unobserved.

Blending inference with selection and explanation

Although inference and selection have historically been studied separately, recently some authors have studied inference and counterfactual simulation together. Gerstenberg et al. (2018) show 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) developed a model of how people attribute mental states to explain behavior; their model holds that people attribute a belief to an agent when the posterior probability that the agent holds this belief is high, and the belief is counterfactually relevant for the observed behavior.

Other researchers have approached how people make inferences *from* an explanation by deploying counterfactual thinking (Kirfel et al., 2022; Nam et al., 2023; Navarre et al., 2024). 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, and infer the causal structure if they know the events' normality. Nam et al. (2023) and Navarre et al. (2024) found people can learn complicated causal structures from token-level explanations.

Explanations as sharing computations

It has been suggested that successful communication and coordination between two people relies on both taking a share of the computational "work" (Christian & Griffiths, 2016).

These authors conceptualized a thinker as performing "computational kindness" for another person when narrowing down the available options in a complex scenario. To apply this idea to explanation: the explainer may obtain information by performing some computation, and then share the results of their work. Part of the function of an explanation would thus be to package the results of computation for transfer to another person. Some results may be more substantial or valuable than others. In sum, information content could contribute to the intuitive appeal of an explanation (Ying et al., 2025).

The current work

In this paper, we study causal explanation in a context where some events are unobserved. For example, in one scenario, a company wins a contract if they give a presentation on either of two new features, provided the files underlying the features work. Say we can see the salesperson presented both features and won the contract, and we know how often they present each feature and how often the files malfunction, but we don't know which feature(s) impressed the client. Why do participants think the contract was won? This task requires causal selection (because there are four potential causes) as well as inference (because two of them are unobserved).

We outline a computational framework for causal explanation in the presence of unobserved variables, and report results of an experiment testing the predictions of this model.

Computational framework

Our computational model has several components which we introduce verbally and in a general mathematical framework, which we then apply in more detail to our specific situation. Building up a model from components means we can switch them in and out later to assess their relative contribution. In addition to causal inference and causal selection, we also include actual causation and computational kindness. Our goal is to assign an overall causal score $\mathcal{C}(X = x)$ to each event, such that an event with a higher causal score is a better candidate for a causal explanation. To outline, the overall causal score comprises:

- **Causal selection.** For this we use *Counterfactual Effect Size* (CES; Quillien & Lucas,

2023) but in principle other models of causal selection could be substituted (e.g. Lagnado et al., 2013; Icard et al., 2017).

- **Actual causation.** The CES model does not automatically weed out non-causes, so we need a step of processing that assigns a score of 0 to variable states that did not causally influence the outcome at all.

- **Inference.** We update the probability of the unobserved variables by conditioning on the observed variables, using Bayes theorem.

- **Computational kindness.** We use information gain as a measure of computational work.

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 in the following sections.

We assume the reasoner knows the causal structure of the relevant system. We operationalize this structure using the formalism of Structural Causal Models (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 defined as functions of their parents, i.e., *structural equations* describe the causal relationships between the variables. The upshot of using functions is that the specified causal relationships of $\mathcal{V}$ are deterministic, and any stochasticity in the system comes from $\mathcal{U}$. Variables labeled with capital letters represent whether each event occurs (for example $C = 1$ means that event C happened and $C = 0$ means it did not happen).

We consider a structure where the effect of a cause node X depends on an unobserved (latent) noise term X_u. 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 can have an effect on E; see Figure 1 for a simplified schema of one node pair.

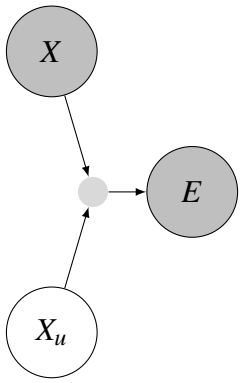

Figure 1

A single 'X-pair' pair of variables: X is drawn from $\mathcal{V}$ and X_u is drawn from $\mathcal{U}$. The small gray joining node is our innovation to show variables occur in conjunctive pairs.

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 $\mathbf{V}$ the set of variables other than E and X. We will also write $S(X = x|\mathbf{V}{=}\mathbf{v})$ to express the selection score that would be assigned to $X = x$ under the assumption that $\mathbf{V}{=}\mathbf{v}$ 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 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, 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 exogenous variable from a probability distribution, and then setting the effect variables according to their structural equations. When simulating a possible counterfactual world, each exogenous variable V is sampled from the probability distribution $s\delta(V) + (1-s)P(V)$, where $\delta(V)$ is the value of V in the actual world, $P(V)$ is the prior probability of V, and s is a 'stability' parameter governing the degree of anchoring to what happened in the real world (Lucas & Kemp, 2015; Quillien et al., 2023).

The CES score of $X = x$ for $E = e$ is then computed on the basis of the simulated possibilities. In our setting, the score computed by the model is equivalent to the Pearson correlation coefficient between $X = x$ and $E = e$ across the simulated counterfactual possibilities.

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 the variable state we are computing it for, and then compute a weighted average of these scores, where the weights are the probabilities of the states of the world.

The steps are as follows: First calculate CES scores, S, for each possible combination of variables, and normalize by e.g., softmax (to enable comparison across models) to obtain $\hat{S}$.

Then the marginal causal selection score S' of $X = x$ is computed 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:

$$S'(X = x) = \sum_{v \in V} \hat{S}(X = x | \mathbf{V} = \mathbf{v}, X = x) P_\alpha(\mathbf{v} | X = x) \tag{3}$$

Computing the total 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, 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, normalized CES score S' by the posterior probability of the variable value, $X = x$.

$$\tilde{S}(X = x) = S'(X = x) P_\alpha(X = x) \tag{4}$$

Measuring the value of computation: information gain

Explanations are informative; the act of explaining is inherently about increasing and transferring information. We think it helpful to recruit information theory — the mathematical study of communicating information and measuring uncertainty — to the arena of causal explanation, especially in our current context of uncertainty. Specifically, we wonder whether explainers may prefer to give explanations with high information content (see also Ying et al., 2025; Klopfenstein & Mercier, 2026).

In the previous section the explainer obtained the posterior probability of a variable. This probability update carries some computational cost. We thus need a way to measure the value of the posterior probability and allow it to scale: we need a proxy for the *value of computation*.

We use Information Gain ('IG') as a measure of distance between prior and posterior of unobserved variables to represent the amount of computation or cognitive 'kindness' explainers expend when inferring the value of a variable.

Information gain is defined as

$$\mathrm{IG}(X_u) = H(P_\alpha(X_u)) - H(P(X_u)), \tag{5}$$

where H denotes the Shannon entropy

$$H(X) = -\sum_{x \in X} P(x) \log_2 P(x). \tag{6}$$

By definition information gain is zero for any observed variable because its prior and posterior entropy are zero. Therefore, the intuition behind this component is that it increases the selection value of mentions of variables that have been inferred rather than given.

We now define the information gain for an unobserved variable as quantity K:

$$K = \mathrm{IG}(X_u) \tag{7}$$

This quantity K extends the expected causal score given in (4) and its importance is scaled by a parameter κ. The overall causal score $\mathcal{C}(X = x)$ is then computed by scaling the value of the inference over the inferred variables and checking for actual causation.

$$\mathcal{C}(X = x) = (\tilde{S}(X = x) + \kappa K) T(X = x) \tag{8}$$

where $T(X = x)$ is 1 if $X = x$ is an actual cause of E, and 0 otherwise.

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

Singular causal explanation with two X-pairs

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. 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:

$$E := (A \wedge A_u) \wedge (B \wedge B_u) \tag{9}$$

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

$$E := (A \wedge A_u) \vee (B \wedge B_u) \tag{10}$$

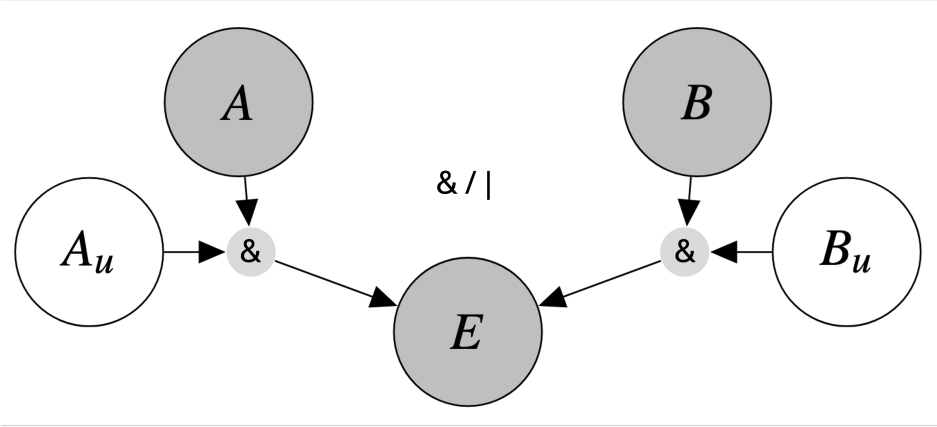

Figure 2

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

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 happened, the reasoner should i) check what events contributed to the outcome (actual causation), ii) engage in causal selection, iii) infer the posterior probability of the latent variables given what they have observed: $\mathrm{P}(A_u, B_u | A, B, E)$, iv) integrate these processes. We discuss each component in turn. For simplicity we only discuss the information gain component of our model in the General Mathematical Framework.

Actual causation

In our computational modeling we use a heuristic to exclude events that do not qualify as actual causes. Specifically, we assign 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 (2005). We also assign a causal score of $\mathcal{C}(X) = 0$ to unobserved variables if their observed 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. Similarly, if E happens and $B = 1$ but $B_u = 0$, then intuitively neither $B = 1$ nor $B_u = 0$ 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 (2005); 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. In this paper we sample 100,000 counterfactuals for each world and for stability we use $s = .7$ on the basis of past model fitting (Lucas & Kemp, 2015; Quillien & Lucas, 2023).

For this first step we postpone treatment of the unobserved variables until the next step. We compute the CES scores, S, for each variable as if we knew the actual world-values of each variable.

For each combination of all variables we then apply a softmax over the causal selection scores of the four variables, with temperature $\tau 1$ to yield a normalized selection score, $\hat{S}$.

$$\hat{S}(X = x) = \frac{\exp(\mathcal{S}(x)/\tau 1)}{\sum_{x \in N} \exp(\mathcal{S}(x)/\tau 1)} \tag{11}$$

where in our case N is always four because each of A, A_u, B, B_u always has a value.

Causal inference

We assume the explainer infers the values of A_u and B_u using Bayes' rule:

$$P(A_u, B_u | A, B, E) = \frac{P(E|A_u, B_u, A, B)P(A_u, B_u)}{P(E|A, B)} \tag{12}$$

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

We now assume the explainer marginalizes over the possible values of the two unobserved variables A_u and B_u, using a specific form of the general Equation 3. For example, to compute the marginal selection score for $A = a$, denoted $S'(A = a)$, we compute:

$$S'(A=a) = \sum_{A_u,B_u} \hat{S}(A=a|A=a,B=b,A_u,B_u)P_\alpha(A_u,B_u) \tag{13}$$

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

Next we compute an *expected* counterfactual effect size $\tilde{S}$ by multiplying the marginal selection score S' by the posterior probability of the variable value as in Equation 4. For example for $A_u = 1$:

$$\begin{aligned}
\tilde{S}(A_u=1) &= S'(A_u=1)P_\alpha(A_u=1) && (14)\\
&= \sum_{B_u} S'(A_u=1|A=a,B=b,A_u=1,B_u) \\
&\quad \times P_\alpha(B_u|A_u=1)P_\alpha(A_u=1) && (15)\\
&= \sum_{B_u} S'(A_u=1|A=a,B=b,A_u=1,B_u) \\
&\quad \times P_\alpha(A_u=1,B_u) && (16)
\end{aligned}$$

Combining the four cognitive modules

We assume the four elements of selection, inference, actual causation and computational kindness are combined as in Equation 8.

Choice model

The sections above specify how the model assigns causal scores to variables. To convert these causal scores to predicted choice proportions over our set of eight categorical choices, we assume that participants are mixing random choices with choices that soft-max over the causal scores:

$$P(\text{choice}=x) = \frac{\varepsilon}{|N|} + (1-\varepsilon)\frac{\exp(C(x)/\tau 2)}{\sum_{x\in N}\exp(C(x)/\tau 2)} \tag{17}$$

N is the set of variable states participants select from (from $A=0$; $A=1$; $A_u=0$; $A_u=1$; $B=0$; $B=1$; $B_u=0$; $B_u=1$), $\tau 2$ is a temperature parameter (higher values indicate more

stochasticity) and ε is a random noise parameter (to handle any erroneous responses or others impossible under the model, chiefly the values A and B did not take), and $C(X)$ is the composite causal score.

Lesioned models

We will also explore 'lesioned' models to assess our claim that when people make a causal judgment, they engage both in inference (about the value of unobserved causes) and in causal selection. Because we also investigate the role of actual causation and computational kindness, this gives four 'modules' which can be switched in and out relatively independently (except that the kindness module cannot occur without the inference module to act upon). We hypothesized all four to be important to a full account of how people provide causal explanations but test this with our model fitting. 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

In the full model and those with the actual causation module, variable states that cannot be an actual cause are set to 0 ($-\infty$ for optimization). We also test variants of the models defined above that *do not* check if an event is an actual cause of the outcome. In practice this means causal scores are calculated for all variables following the full computational framework, and none are then removed.

Lesioning causal selection

To lesion the causal selection module, we assume that people do not engage in counterfactual simulation when making causal judgments. Once they determine which variables are actual causes of E, they select these variables simply in function of their posterior probabilities. In terms of the mathematical framework defined above, we replace all CES scores S with 1.

Lesioning inference

To remove the inference module, instead of setting $P_\alpha(A_u, B_u)$ to be 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, but note the `no Inference` models are functionally equivalent to `no Kindness; no Inference` models.

Lesioning both inference and selection

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

Lesioning kindness

These models assume that people do not favor explanations that have high information gain. The term K is set to 0 and otherwise the series of models works as above.

Experiment

Methods

We conducted a behavioral experiment to test our models.[1]

Design

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 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[2] and Appendix A)

[1] Link removed for review

[2] Five worlds involve the conjunctive structure defined in Equation 9, and seven involve the disjunctive structure defined in Equation 10. Unequal split is due to the fact some observation patterns are possible for the disjunctive but not conjunctive structure (e.g., $A = 1$, $B = 0$, $E = 1$).

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. 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.

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.

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)

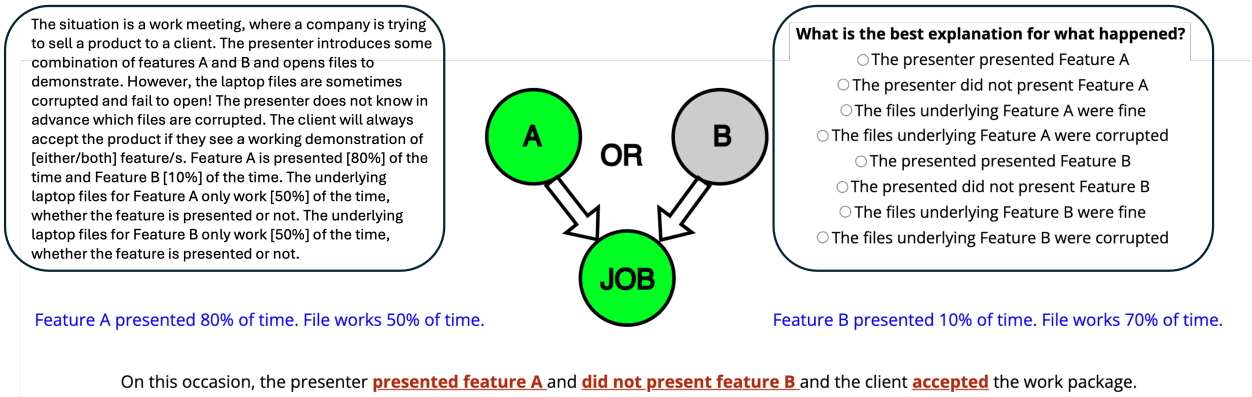

Figure 3

Simplified schematic of one trial: blue text gives base rates; grey/red text and graph describe what happened this time.

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.

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.

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. 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 Anonymised Repository.

Results

A Fisher's exact test found no statistical difference between the three cover stories, for each of the 36 worlds[3] and so our analyses collapse across cover stories.

Figure 4 shows the choice proportions of participants in probability setting 3 [$P(A) = .1, P(A_u) = .7, P(B) = .8, P(B_u) = .5$]. Figures showing full results in probability settings 1 and 2 are available in Appendix C.

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-adjusted for three comparisons, Cramer's $V = .215$.[4]

In the following subsections we discuss the models. First we fit the series of models in aggregate and by participant, and then comment on the best fitting models and the relative contribution of each of the four modules of causal cognition.

---

[3] 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

[4] Two rows containing zeroes were removed from the 96 (8 choices * 12 scenarios) so the test was over three vectors of length 94.

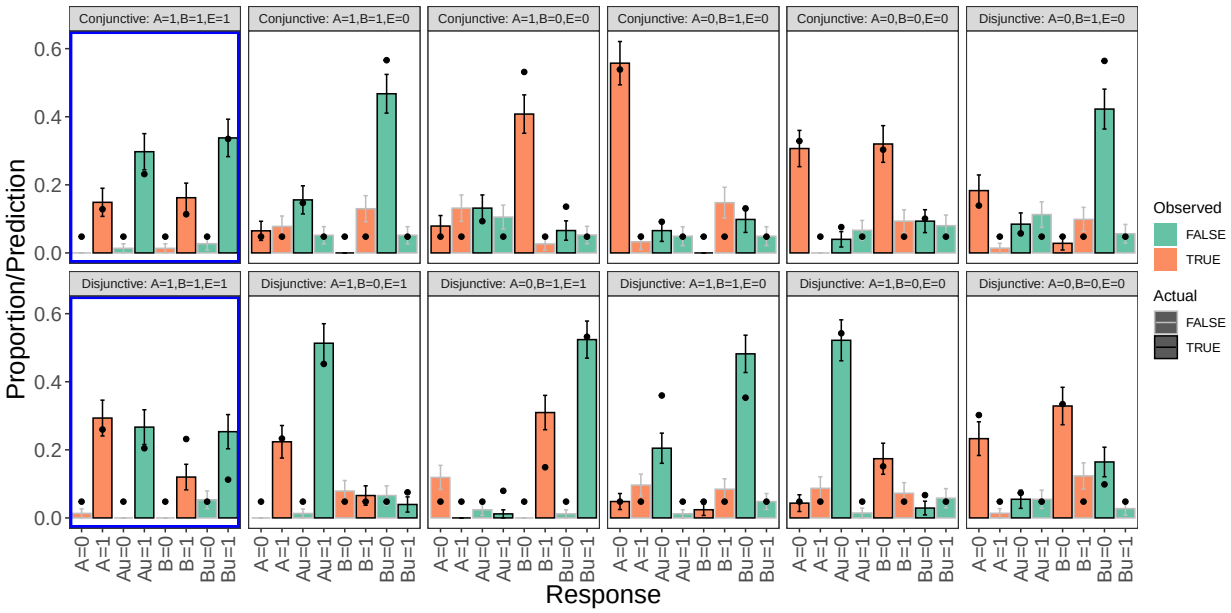

Figure 4

Results in Setting 3 $P(A) = .1, P(A_u) = .7, P(B) = .8, P(B_u) = .5$. *Participants (bars) plus* full *model with fitted* ε, κ *and* τ *(black dots). Blue highlights for canonical "everything happened" scenario (*$A = 1$, $B = 1$, $E = 1$. *See Repository for plots of the other probability conditions.*

Model fit

We fit the models to the data by minimizing negative log likelihood, with two softmax temperatures τ1 and τ2, the noise ε, and the information gain 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 module is signified by 'no-`X`' in the name.[5]

The item-level Pearson correlation coefficient between the full model and participants' average judgments was $r(286) = .904$, $p < .001^{***}$. Model predictions for the full model (for one

[5] 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.

of the probability settings) are shown as black dots on Figure 4. See also Appendix C for results in other probability settings.

Table 3

Parameter Values and Performance Metrics for All Models

	Model	τ1	ε	τ2	κ	logl	BIC	N participants
1	Full	5.878	.383	.0791	.218	-4251	8533	42
2	**no ActualCause**	**2.107**	**.206**	**.0903**	**.177**	**-4222**	**8476**	**23**
3	no Selection	2.718	.384	.0736	.221	-4255	8542	5
4	no ActualCause; no Selection	2.718	.211	.0797	.265	-4300	8631	0
5	no Kindness	2.663	.383	.0785	-	-4325	8673	39
6	no ActualCause; no Kindness	1.45	.212	.0971	-	-4276	8577	29
7	no Inference; no Kindness	.382	.382	.5132	-	-4545	9114	12
8	no Kindness; no Selection	2.718	.389	.0629	-	-4358	8739	1
9	no ActualCause; no Inference; no Kindness	.348	.188	.2911	-	-4673	9370	22
10	no ActualCause; no Kindness; no Selection	2.718	.223	.0795	-	-4520	9064	1
11	no Inference; no Kindness; no Selection	2.718	.383	.3817	-	-4581	9186	0
12	no ActualCause; no Inference; no Kindness; no Selection	2.718	.205	.3219	-	-4956	9936	2
13	Baseline					-5365	10730	39

Note. Model with the lowest Bayesian Information Criterion in bold. 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 the same as the baseline which assigns uniform probability to all answers.

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 summed up 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: $\chi^2 = 198.3$, df = 12, $p < 2.2e-16^{***}$. The fits largely follow the other fit metrics: the `full` model and the model lesioned to have no information gain (`no Kindness`) fit the highest numbers of participants, and the `no Actual` model also fit large numbers of participants, although 18.1% of participants were best fit by the uniform baseline.

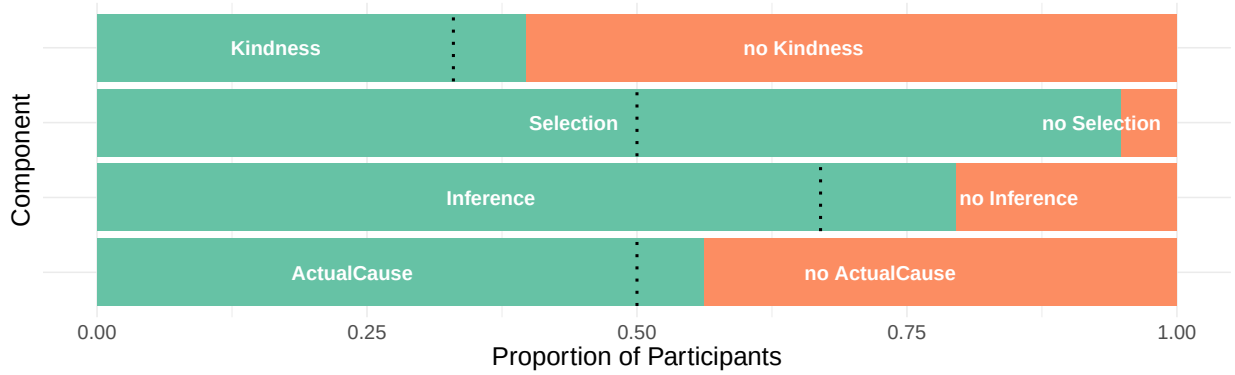

Figure 5

Proportions of participants best fit by models with and without the four modules (excluding the 39 best fit by a uniform baseline). Dotted lines are the proportions expected if participants were allocated to models uniformly.

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 the uniform baseline model), and then summed the number of participants best fit by the models in each bucket. This brings out the relative differences (results shown in Figure 5). We compared the actual numbers of participants to the expected

proportions.[6] We ran a simple χ^2 for each module for whether the counts in each bucket 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 but the difference was not significant (56.25% where random chance would be 50%, $\chi^2 = 2.75$, df = 1, $p = .097$). More participants were fit by models containing the `Inference` module than expected under uniform distribution (79.5% where random chance would be 67.8%, $\chi^2 = 13.08$, df = 1, $p = .0003^{***}$). Far more participants were fit by models containing the `Selection` module than expected under uniform distribution (94.9% where random chance would be 50%, $\chi^2 = 141.8$, df = 1, $p < 2.2e-16^{***}$). More participants were fit by models containing the `Kindness` module than expected but the difference was not significant (39.8% where random chance would be 33.3%, $\chi^2 = 3.32$, df = 1, $p = .07$).

Summary of results

Our full model fit well on BIC, suggesting all four modeled modules were used to some degree in our causal explanation task.

We now drill down in more detail to discuss each module. 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` cognitive module. Each model containing it fit better on BIC than the equivalent model without it (for example, `no Kindness` fit better than `no Kindness; no Selection`, etc.). Far more participants were fit by models containing the `Selection` cognitive module than without.

---

[6] 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.

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 Inference; no Kindness` fit better than `no Inference; no Kindness; no Selection`, etc.). Far more participants were fit by models containing the `Inference` cognitive module than without.

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.). More participants were fit by models containing the `Inference` cognitive module than without (although this difference is not significant).

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 does improve fit: for example, `no Inference; no Kindness` fit better than `no ActualCause; no Inference; no Kindness`, etc. More participants were fit by models containing the `ActualCause` cognitive module than without (although this difference is not significant).

Both the model fitting and the by-participant fitting suggest actual causality contributes less to causal explanation than the other modules. This is consistent with our lesser focus on parametrizing actual causality: we incorporated it more from necessity in implementation rather than attempting an in-depth theory. The best-fitting model was narrowly the one lesioned to turn off the actual causation module: that is, it gave causal score to variables taking values they did not currently take. It is likely some participants ended up selecting a non-actual cause because of inattention, even after investing time and resources in computation (that is, they were still influenced by the value of the posterior and CES scores). This would explain why this erroneous selection of non-actual causes is not entirely modeled by the noise parameter ε, which gives a uniform score over all variable values.

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 combine them. We considered how people give causal explanations in a setting where some events are unobserved, such that people need to engage in both causal inference and causal selection to reason about the cause of what is observed. First, we sketched a computational framework for how these two processes can be integrated. We expanded this computational framework with the concept of information gain from information theory, to allow us to further quantify the contribution of causal inference to reducing uncertainty. Then we reported the results of an experiment testing how people give causal explanations under this uncertainty, and showed by comparison with predictions of our model that causal explanation likely involves causal selection, causal inference and information gain.

Our experimental data 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) in the way a significant proportion of our participants were best fit by models containing causal selection, and few were best fit by models without it. Our results provide a timely extension to the existing literature on this model (Konuk et al., 2023; Quillien & Barlev, 2022) by showing it can account for people's explanatory behavior in richer settings containing some uncertainty through unobserved variables.

Our experimental data suggests people engage in inference, where they must perform

computations in accordance with probability theory to impute the possible values of what they cannot directly observe and then assess the relative priority of each possible value depending on how likely it is. We modeled this as the human mind performing Bayesian updating on the given prior probabilities. In finer detail, our work continues the tradition of Kirfel et al. (2022) and others who study how the acts of inference and explanation interact. In addition to the Bayesian element as such, our work continues the unification of inference and causality found in causal cognition research, and supports specifically causal model theory as a productive framework within which to study explanatory behavior.

Our primary innovation led to our primary findings: that people not only incorporate inference over the value of unobserved variables, but scales the value of the information obtained by performing that inference. We posited there was a role for parameterizing people's willingness to do some thinking on behalf of the other, making explanations citing inferred states more valuable than those citing observed states, other things being equal. We conceptualized this as a selection bonus for variables that scaled with information gain from prior to posterior. We find that participants' explanation choices were indeed sensitive to information gain, suggesting that people favor explanations that provides information about unobserved events.

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 has modules for causal inference, causal selection, and information gain, and found that each of these modules was necessary to account for people's explanatory preferences. Our work contributes to painting a fuller picture of how people decide what makes for a good explanation under conditions of uncertainty.

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., 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.

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.

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.

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. (2005). Causes and explanations: A structural-model approach. part i: Causes. *The British journal for the philosophy of science*.

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

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

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.

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

Klopfenstein, A., & Mercier, H. (2026). Explaining is not enough: Appealing explanations should also be surprising. *Psychonomic Bulletin & Review*, *33*(1), 35.

Konuk, C., Goodale, M. E., Quillien, T., & Mascarenhas, S. (2023). Plural causes in causal judgment. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45).

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).

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).

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

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*.

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. Existence of columnar gray banding signifies no meaning other than to assist visual clarity. *Width* of each gray band, however, indicates number of possible settings of unobserved variables: for example, in d6 and c5 both unobserved variables have only one possible value and in d2:d5 one unobserved variable is known but the other could take either value.

Appendix B

Cover stories

The behavioral experiment was presented through three cover stories, shown below. Here they are presented in slightly condensed form because, firstly, in the actual experiment they were shown on click through screens with accompanying figures and, secondly, certain phrases could change depending on the situation. Here that is shown in square brackets but in the real experiment it was incorporated into the text. 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

Full model in probability settings 1 and 2

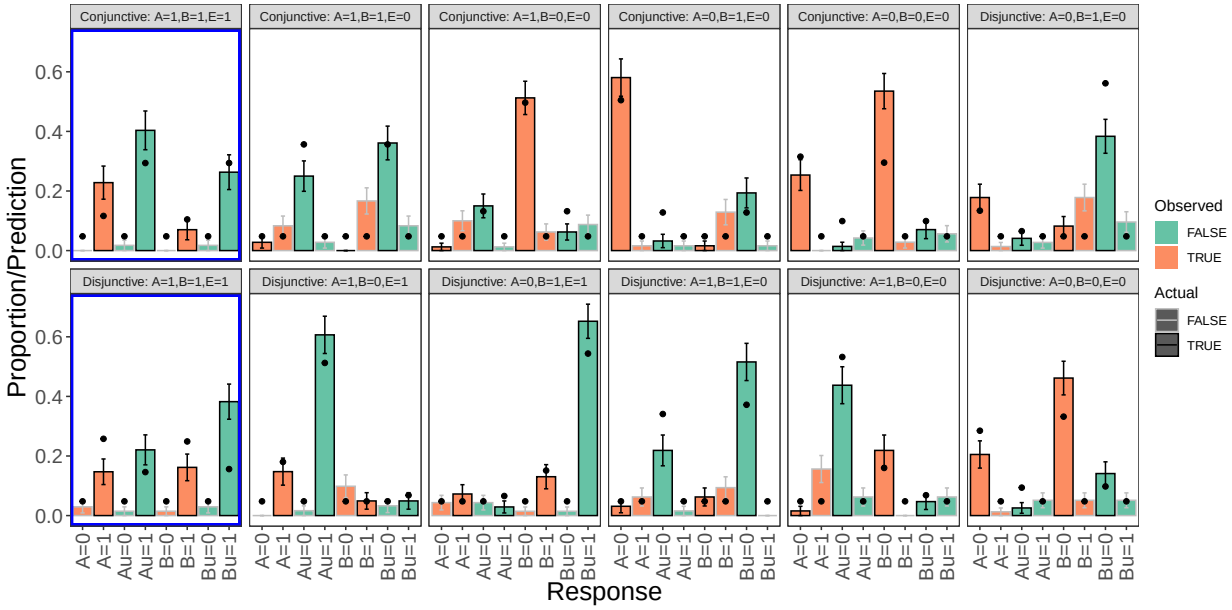

Figure C1

Full model (dots) on participant (bars), probability setting 1

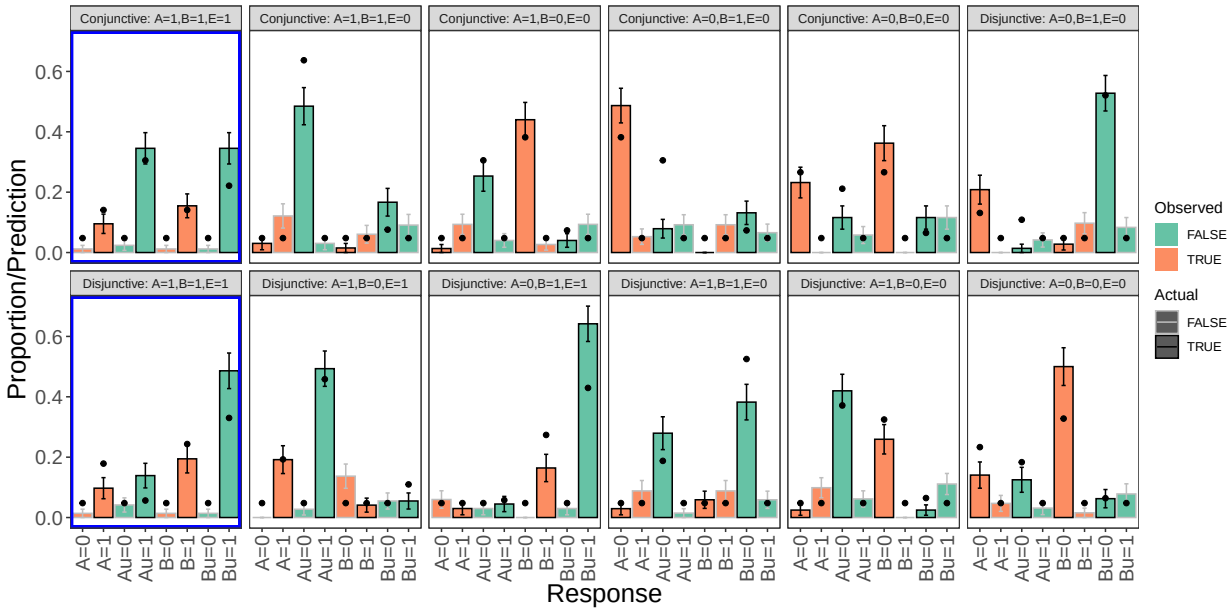

Figure C2

Full model (dots) on participant (bars), probability setting 2