# Show and tell: Learning causal structures from observations and explanations

**Andrew Nam** (andrewnam@stanford.edu), Department of Psychology, Stanford University

**Christopher Hughes** (chughes4@stanford.edu), Department of Philosophy, Stanford University

**Thomas Icard** (icard@stanford.edu), Department of Philosophy, Stanford University

**Tobias Gerstenberg** (gerstenberg@stanford.edu), Department of Psychology, Stanford University

## Abstract

There are at least three ways of learning how the world works: learning from observations, from interventions, and from explanations. Prior work on causal inference focused on how people learn causal structures through observation and intervention. Our study is the first to look at how explanations support causal structure learning. We develop a normative inference model that learns from observations and explanations, and compare the model's predictions to participants' judgments. The task is to infer the causal connections in 3-node graphs, based on information about their co-activation, and explanations of the kind "B activated because A activated". We find that participants learn better from explanations than from observations. However, while the normative model benefits from having observations in addition to explanations, participants did not.

**Keywords**: causality; explanation; counterfactuals; learning; inference.

## Introduction

How do people figure out how the world works? Imagine that your friend Kingsley gifts you a brand new 'Cor 10,000' for your birthday. You're thrilled but you also have no idea how this fancy-looking device works. Kingsley challenges you to figure it out. It has three components with flashy lights on top, labeled A, B, and C. When a component activates the light turns on. To help you out a little, Kingsley tells you that some components make the others go. So, for example, it's possible that component A activates component B, or that B activates A, or that there is no direct connection between them. Your task is to learn which components activate which.

There are at least three ways of going about this. First, you could just observe the device. The pattern of statistical regularities provide some evidence about the underlying causal model. If A and B repeatedly activate, but not C, it is plausible that there is a connection between A and B (and no connection with C). However, you would not know whether A activates B, or vice versa. Second, you could take some action by turning A on and seeing whether B also turns on. If so, this would suggest that A activates B and not the other way around. Third, you could ask for help. As it turns out, the 'Cor 10,000' comes with a little semi-helpful robot gives an explanation every time something happens. 'Semi-helpful' because while what the robot says is always true, it doesn't necessarily give the most informative explanations. For example, when A and C activated but not B, the robot might tell you "C activated, but not because A activated". Given this explanation, it is still possible that A and C are not connected to one another and that both activated by themselves, or that it is in fact C that activates A.

In this paper, we study how people combine information from observations and explanations in causal inference. We first briefly review prior work and motivate our setup. We then describe a normative computational model that learns from observations and explanations. We compare the model's predictions with human causal inferences in an experiment that contrasts learning from observations, learning from explanations, and learning from both. We discuss the implications of this work and point out directions for future research.

### Learning from observations

Most current AI models learn about the world through passive observation. ChatGPT, for example, learned from "observing" large amounts of text from the internet. However, there are limits to what one can learn about the causal structure of the world from observations alone: correlation does not imply causation (but see Kıcıman, Ness, Sharma, & Tan, 2023). Under certain conditions, however, people can learn causal structures from purely observational data, such as when the underlying system is deterministic (Rothe, Deverett, Mayrhofer, & Kemp, 2018), when there is information about the temporal dynamics of the system (Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018), or when inferring physical dynamics (Ullman, Stuhlmüller, Goodman, & Tenenbaum, 2018).

### Learning from actions

Another mode of learning about the world is through active exploration. With only covariational information, it would be impossible to tell apart whether A causes B, or B causes A. Certain causal models, such as $A \rightarrow B$ and $A \leftarrow B$, are 'Markov equivalent' and imply the same statistical dependencies. However, if we could actively intervene on the causal system of interest, we could tease the two models apart. If $A \rightarrow B$ were true, then intervening on A should activate B, whereas if $A \leftarrow B$ were true, intervening on A would not activate B (although B may activate by itself).

Psychologists have studied how people infer the underlying causal structure through passive observation and active intervention (Bramley, Dayan, Griffiths, & Lagnado, 2017; Bramley, Lagnado, & Speekenbrink, 2015; Coenen, Rehder, & Gureckis, 2015; Gong, Gerstenberg, Mayrhofer, & Bramley,



In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.* 

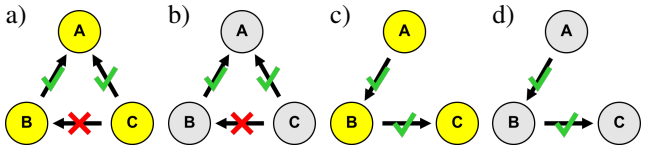


Figure 1: Illustration of different causal devices and events. Yellow nodes are active and gray nodes are not. Connections with a ✓ worked, and connections with an ✗ didn't work. In a), the activation of A is causally overdetermined. In c) there is a causal chain of activations from A to B, and B to C.

2023; Meder, Gerstenberg, Hagmayer, & Waldmann, 2010; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003; Waldmann & Hagmayer, 2005). Generally, this work has found that actively intervening helps with learning, and that people make better causal judgments when they choose interventions themselves, rather than observing someone else intervening on the world (e.g. Bramley et al., 2015).

### Learning from explanations

A third mode of learning about the world is by receiving explanations from others (Keil, 2006; Lombrozo & Vasilyeva, 2017). "B happened because of A" not only tells you that A and B both happened, but also that there exists a causal relationship between A and B. One way to analyze the meaning of "because" in causal explanations is by way of a counterfactual. Accordingly, "because" means that the following counterfactual is true: B would not have happened if A had not happened. Moreover, because people share systematic preferences about how to explain things (Hilton, 1990), they can infer much more from explanations than what is explicitly communicated (Kirfel, Icard, & Gerstenberg, 2022).

### Learning from observations and explanations

These three modes of learning about the world map roughly onto the three levels of the causal hierarchy proposed by Pearl (2000) (see also Bareinboim, Correa, Ibeling, & Icard, 2022; Gerstenberg, 2022). On level I, we can answer questions about statistical dependence of the form $p(y|x)$. On level II, we can answer questions about the hypothetical consequences of acting on the world, such as how likely $y$ would happen if we were to intervene on $x$, $p(y|\text{do}(x))$. On level III, we can answer questions about counterfactuals, such as whether $y'$ would have happened if $x'$ had happened, when in fact both $x$ and $y$ happened, $p(y'_{x'}|x,y)$. It is only on this third level of the hierarchy that we can reason about *why* something happened.

Here, we ask the question of how well people can learn about the causal structure of a system from explanations. While prior work has looked at causal structure learning based on observations and interventions, our study is the first to look at learning from observations and explanations.

## Computational model

Our computational model defines a generative process for activations in directed acyclic graphs (DAGs) with 3 nodes and up to 3 edges. It also defines a process for generating explanations. We describe each in turn.

### Generating activations

Figure 1 illustrates some causal devices in action. In our setup, any node in the model without incoming connections has a 50% probability of activating on its own. Nodes with incoming edges have a 10% probability of activating by themselves. Edges have a 90% probability of working. When a parent node activates and an outgoing edge works, the connected child also activates. One active parent node is sufficient to activate a child node with multiple incoming edges. For example, in Figure 1a, component C had a 50% chance of activating. The connection from C to B didn't work while the other two connections worked. B only had a 10% chance of activating spontaneously because it has an incoming connection. A activated here because both B and C activated. A would have also activated if only B (or only C) had activated.

There are 25 possible 3-node DAGs. We classify these into 6 different types of causal networks (see Figure 3a) where all graphs in each category are isomorphs.

### Generating explanations

For a given causal network and pattern of activations, our model generates various possible explanations. We use the following four templates for generating explanations:

1. **factual**$^+$: "Y activated because X activated" (e.g., A activated because B activated in Figure 1a).
2. **factual**$^-$: "Y activated, but not because X activated" (e.g., B activated but not because C activated in Figure 1b).
3. **counterfactual**$^+$: "Y would have activated if X had activated" (e.g., B would have activated if A had activated in Figure 1c).
4. **counterfactual**$^-$: "Y would not have activated, even if X had activated" (e.g., A would not have activated, even if C had activated in Figure 1d).

Explanations describe two components that either both activated or did not activate. As alluded to earlier, we adopt a counterfactual semantics of "because". Accordingly, "Y activated because of X" is true if both X and Y activated, and if it is true that Y would not have activated had X not activated. Similarly, "Y activated but not because X activated" is true when Y would have activated even if X hadn't activated.

If a child component self-activates, its connection to parent components are considered inactive, such as the connection from C to B in Figure 1a. While explanations of the type factual$^-$ may be less common in everyday life, we included them in our study because we were interested to see what inferences people draw from them.

**Overdetermination** A simple counterfactual definition of "because" fails in situations of overdetermination such as in Figures 1a and 1b. To handle such situations, we use the model of actual causation by Halpern and Pearl (2005) which uses a more relaxed criterion of counterfactual dependence.

An event can qualify as a cause of an outcome even when there is no counterfactual dependence in the actual situation, as long as there is an admissible contingency in which the counterfactual would have been true. In Figure 1a, "B caused A to activate" is true because in the event that C had not been active, A would have been counterfactually dependent on B.

**Causal chains** Figure 1c depicts a causal chain. Here, the explanation "C activated because A activated" is valid because C would not have activated had A not activated.

## Normative inference model

Given instances of activations and explanations, the goal is to infer the underlying causal graph (i.e. the existence and directionality of edges). For each device $d_i$ in the set of devices $\mathcal{D}$, we start with a uniform prior $P^{(0)}(d_i) = \frac{1}{25}$. Then, after each trial $t$, given observation $o^{(t)}$ and explanation $x^{(t)}$, we update the posterior probability for each device $d_i$ using Bayes' rule. We define the union of node and edge activations as an *event*. Since the event is only partially observable and multiple events can produce the same observations (e.g. Figures 1a and 1c), we marginalize over the possible events when performing the update using the conditional probability of the event given the device $P(e|d_i)$. We denote the set of events consistent with the given observation $o$ and explanation $x$ using $\mathcal{E}(o,x)$. An explanation is sampled from an event, and we represent this using $P(x^{(t)}|e)$. Put together, the full update rule for the normative model is

$$\mathrm{P}^{(t)}(d_i|o^{(t)},x^{(t)}) = \frac{\sum_{e\in\mathcal{E}(o^{(t)},x^{(t)})} P(x^{(t)}|e)P(e|d_i)P^{(t-1)}(d_i)}{\sum_{d\in\mathcal{D}}\sum_{e\in\mathcal{E}(o^{(t)},x^{(t)})} P(x^{(t)}|e)P(e|d)P^{(t-1)}(d)}$$

In the 'observation only' condition, we update $P^{(t)}(d_i|o^{(t)})$ and consider events consistent with only the observation: $\mathcal{E}(o^{(t)})$. Likewise, in the 'explanation only' condition, we update $P^{(t)}(d_i|x^{(t)})$ and consider events consistent with only the explanation: $\mathcal{E}(x^{(t)})$.

To compare the model with participants' responses, we convert the posterior distributions over devices into distributions over connection states. For each connection $c$ in $\{AB, AC, BC\}$, we determine the probability of each state $s$ in {forward, backward, none} by marginalizing out the devices:

$$P(c;s) = \sum_{d\in\mathcal{D}} P(d)\mathbb{1}(\text{connection } c \text{ of device } d \text{ is state } s)$$

# Experiment

In this pre-registered experiment, we compare people's causal structure inferences against the predictions of the normative model. The pre-registration and all materials including the code and data for reproduction can be found at `https://github.com/cicl-stanford/show_and_tell`.

## Methods

**Participants & Design** We collected data via Prolific from 156 US-based participants (*age*: M = 38, SD = 13; *gender*: 87 male, 65 female, 4 no response; *race*: 116 White, 11 Asian, 10 Black, 9 Mixed, 3 Other, 7 no response).

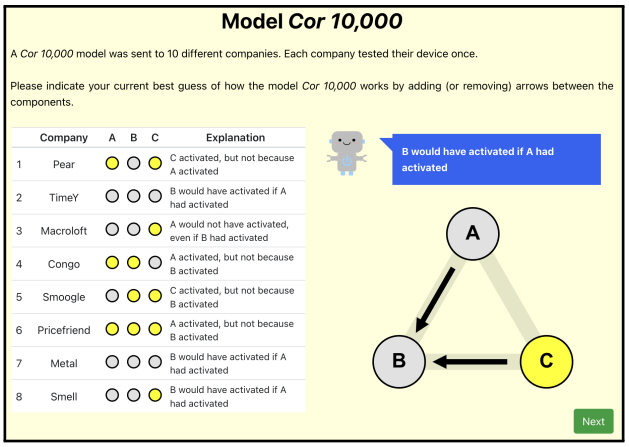


Figure 2: Screenshot of the experiment for 'observation + explanation' condition at trial 8 from a Pair device with connection A → B. In this example, the participant correctly inferred A → B but incorrectly inferred C → B.

50 participants were assigned to the 'observation only' condition, 53 to 'explanation only', and 53 to 'observation + explanation'. Participants were compensated a base rate of $10/hour and a bonus of $0.03 for each edge guessed correctly, with average total earnings of $13.74/hour.

**Procedure** Participants were guided through instructions on using the program interface and how the causal devices work, including information on how they activate, how causal connections work, and what sorts of explanations are valid for different activation events. The task was described using a vignette about companies that are each given a device of the same model, and try to figure out how it works. Participants learned that they would receive explanations from a robot who is truthful but not necessarily the most informative.

To ensure that participants understood everything, they were given a series of comprehension checks throughout the instruction sequence. Before the start of the main experiment, participants were informed that they would receive observations only, explanations only, or both. Figure 2 shows an example of a trial for a participant in the 'observation + explanations' condition. Participants proceeded through 6 sets of 10 trials, where each trial of a set consisted of an event from the same graph. During each trial, participants would be shown an observation, explanation, or both depending on their assigned condition, and could indicate their current belief about how the device works by clicking on the shaded area between the nodes to toggle through arrows in either direction, or no arrow. Participants would also be shown the full history of observations and explanations from previous trials of the same set (depending on their condition). On the first trial of each set, the response interface started with no connections. After the last trial of each set, participants were shown the true underlying graph.

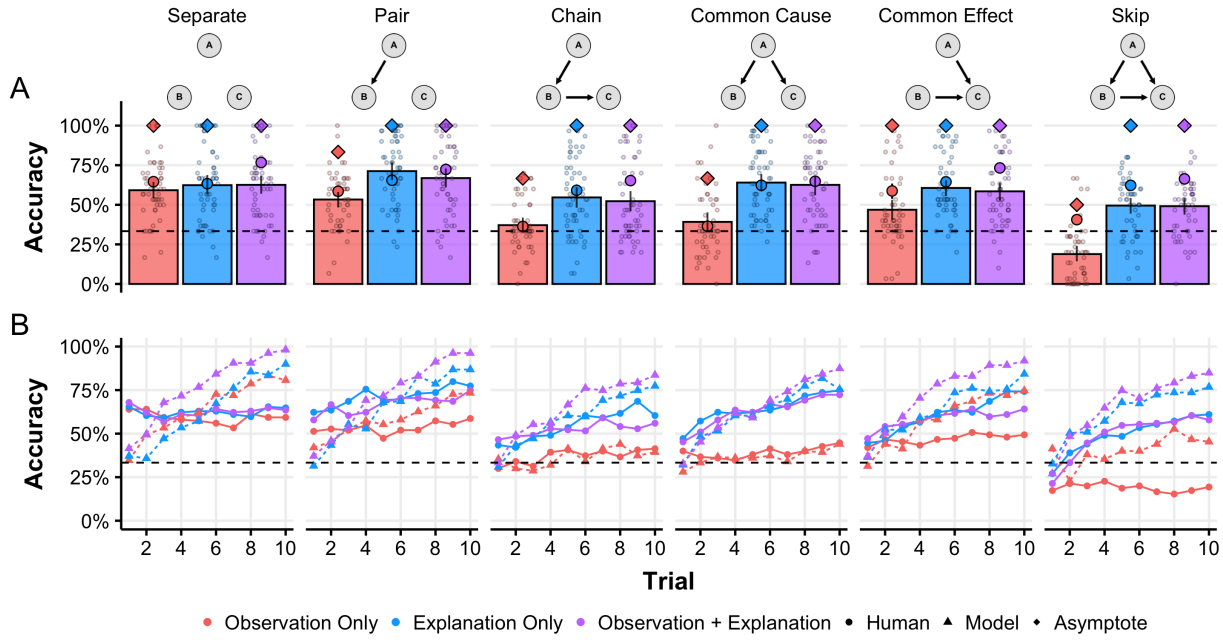


Figure 3: Average accuracy measures for each condition and device type, measured as proportion of edges inferred correctly. Since there are 3 possible edge states, chance accuracy is 33%. **A**. Accuracy over all 10 trials. Bars represent average human accuracy, circles indicate model accuracy, and diamonds indicate asymptotic accuracy (what is theoretically possible given infinite trials). Error bars are 95% bootstrapped confidence intervals. **B**. Average accuracy on each trial.

## Results

**Confirmatory analyses** Figure 3 shows participants' accuracy and model predictions on the different device types, separated by experimental condition. To compare participants' overall accuracy (which we define as the proportion of edge directions correctly inferred) between the three conditions, we use a mixed-effects logistic regression model with pairwise contrasts, fitted to whether or not the participants' responses matched the true edge. We found that participants in the 'observation only' condition had significantly lower average accuracy than in the 'explanation only' condition ($\beta = -0.70, \text{SE} = 0.10, p < .001$; $\beta$ and SE are in logits) and in the 'observation + explanation' condition ($\beta = -0.76$, SE = 0.10, $p < .001$). However, participants' accuracy in the 'explanation only' versus 'observation + explanation' condition did not differ significantly ($\beta = 0.07, \text{SE} = 0.01, p = .507$). This is in contrast to the normative inference model for which having both observations and explanations leads to significantly higher average accuracy than having only explanations ($\beta = 0.33, SE = 0.4, p < .001$), which in turn yields higher accuracy than having only observations ($\beta = 0.51, SE = 0.04, p < .001$). Even when analyzing individual device types, we find that the difference between the 'explanation only' and 'observation + explanation' conditions is not significant for any of the six types (lowest p = .317 for the 'pair' device).

**Exploratory analyses** In the following exploratory analyses, we fitted Bayesian regression models (Bürkner, 2017) to estimate posterior distributions and 95% credible intervals for the predictors.

**'Observation + explanation' vs. 'Explanation only'** One reason for why having both observations and explanations is clearly beneficial compared to having only observations, but not to having only explanations, might be because people make their inferences primarily based on the explanations, even when both types of evidence are offered. We test this hypothesis by applying the normative inference model to the same stimuli presented to participants who received both observations and explanations. For each participant, we apply the model using the data as presented to the participants and, separately, using only the explanations from the same stimuli. We then compute the likelihood of the participants' responses using both models as a measure for how well each model accounts for the data.

Using this method, we find that the model accounts for the data better when using only explanations than when using both observations and explanations for 91% (48 out of 53) of our participants. This suggests that even when participants were offered both observations and explanations, they base their inferences primarily on the explanations.

**How explanations improve inference** Interestingly participants' accuracy in the 'observation only' condition remains relatively flat, as shown in Figure 3b, even in devices where the model exhibits clear increasing trends. We analyze the average accuracy over time using a mixed-effects logistic regression of the form "correct $\sim 1 +$ condition:trial $+$

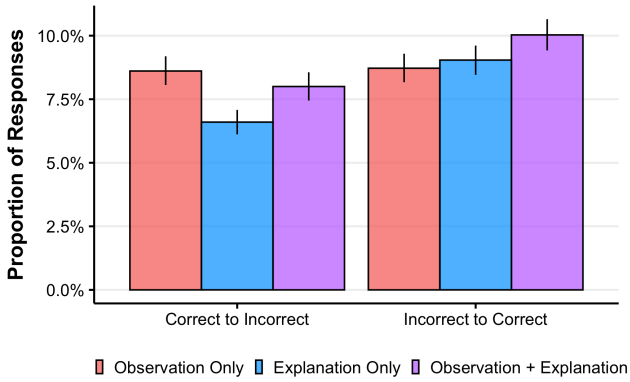


Figure 4: Participants' responses that involved changing a connection from or to a correct state.

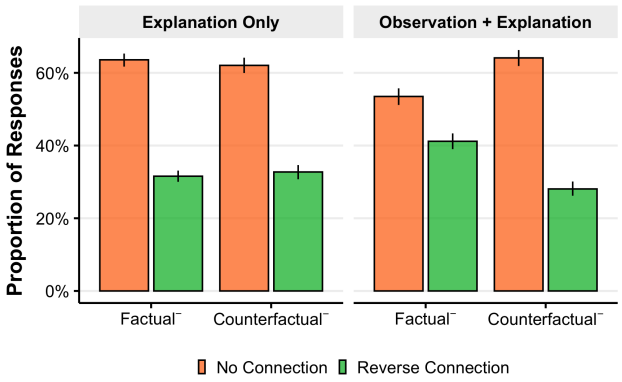


Figure 5: Participants' responses for the connection between X and Y after receiving either a factual$^-$ explanation "Y activated, but not because X activated" or a counterfactual$^-$ explanation "Y would not have activated, even if X had activated". 'Reverse connection' indicates that the connection was set to Y $\rightarrow$ X. 'Forward' responses (X $\rightarrow$ Y) not shown.

(1 | subject)", where "condition:trial" encodes an interaction between condition and trial number. We assign a single intercept 45.7% [26.0, 71.6] for all three conditions since participants share the same degree of uncertainty before having received any information. In the 'observation only' condition, we find that participants' average accuracy does not improve after 10 observations (43.6% [26.7, 55.0]), whereas accuracy improves to 70.6% [50.8, 86.1] after 10 explanations, and to 66.7% [47.0, 86.5] after 10 observations and explanations.

If explanations produce more accurate inferences, what is it about them that helps people to do so? We find that explanations keep participants from mistakenly changing previously correct responses to incorrect ones. Participants in the 'observation only' condition make this switch in 8.6% [8.0, 9.2] of all responses, whereas those in the 'explanation only' and 'observation + explanation' conditions make this switch 6.6% [6.1, 7.1] and 8.0% [7.46, 8.52] of all responses (Figure 4). However, explanations alone are not much more informative than observations alone in helping participants correct previously incorrect responses, though explanations do further improve accuracy when also given observations. We find that on average, participants in the 'observation only' condition make this switch in 8.7% [8.1, 9.3] of all responses, whereas participants in the 'explanation only' and 'observation + explanation' conditions make this switch in 9.0% [8.5, 9.6] and 10.0% [9.4, 10.6] of all responses respectively.

**Interpreting negative explanations** Because different events can produce the same explanations, it is impossible to ascertain the exact device connectivity, especially when the explanation indicates the connection was inactive (*factual*$^-$ and *counterfactual*$^-$ explanations). For instance, if the explanation 'B activated but not because C activated' was given for the event in Figure 1a, one may correctly infer that the connection from C to B did not work in this trial. However, inferring that there is no connection between B and C, or that the connection is from B to C, is also valid and reasonable. How might people handle such ambiguous explanations?

The most direct inference that can be made from a negative explanation is on the very connection referenced in the explanation, such as on the connection between B and C in our above example. As shown in Figure 5, we find that participants typically infer that there is no connection between the two components or that the connection is reversed (i.e. 'B $\rightarrow$ C' in our example). As a point of reference, the normative model's proportion of responses fall between 40.6% and 44.4% for both 'no connection' and 'reverse connection' responses for both factual$^-$ and counterfactual$^-$ explanations.

This bias towards no connections, however, is moderated by whether the participant also sees activations along with the explanations. When only explanations are presented, we find no difference between factual$^-$ explanations and counterfactual$^-$ explanations (31.6% [30.0, 33.3] vs. 32.7% [30 .7, 34.7] for reverse responses). When explanations are presented with observations, however, participants are more likely to respond with reverse connections with factual$^-$ explanations (41.1%, [38.0, 43.2]) than with counterfactual$^-$ explanations (28.1%, [26.1, 30.2]). We note that explanations also provide the same information about the activation states of the two components being described, suggesting that this interaction may be the result of the visual cues that are shown in the 'observation + explanation' condition.

Being told that one component was not the cause of another can also influence inferences on the connections with the third component. Again, assume that what actually happened is shown in Figure 1a, and that a participant received the explanation that "B activated but not because C activated". What inferences do participants draw about the potential connection between A and B, and A and C in this case? First, we note that people and the model don't show notable differences here in their inferences between the 'explanation only' and 'observation + explanation' conditions, and so we aggregate across both conditions for our analyses here.

As shown in Figure 6, both people and the normative model

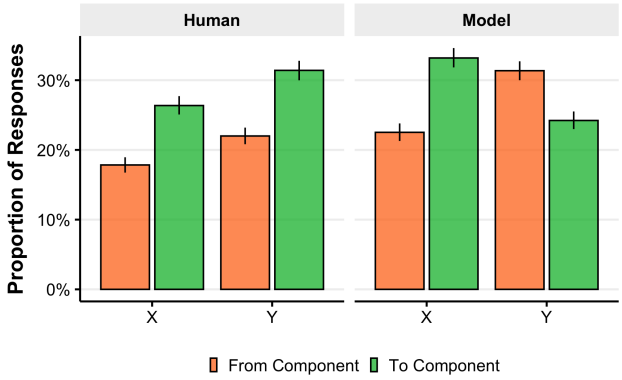


Figure 6: Responses after receiving a factual$^-$ or counterfactual$^-$ explanation of the form "Y activated but not because X activated" or "Y would not have activated, even if X had activated". Responses are for connections between X and Z, and between Y and Z. 'From component' indicates that the arrow points to Z and 'To component' indicates that the arrow points away from Z.

make similar inferences about the connection between A and C. They agree that it is more likely that the $A \leftarrow C$ connection is more likely than $A \rightarrow C$. People and the model disagree, however, about the edge involving the connection between A and B. The model infers that it is more likely that $A \leftarrow B$ (31.3% [30.0, 32.7]) than $A \rightarrow B$ (24.2% [22.9, 25.5]) whereas people infer the opposite (22.0% [20.8, 23.3] vs 31.4% [30.0, 32.8]). When people here that "B activated but not because C activated", they are more likely to infer that there is a connection from $A \rightarrow B$ whereas the model would more is more likely to infer that $A \leftarrow B$.

## General discussion

Explanations play a powerful role in causal inference. They can disentangle merely correlated structures into cause and effect, and accelerate convergence towards the correct underlying structure by providing reliable signals in otherwise highly stochastic environments. In this work, we explored how people use explanations to infer causal structures of simple 3-node DAGs. While explanations help with learning, participants' inferences did not match those of our normative model. Interestingly, people deviated from optimality systematically, suggesting that humans may possess certain biases for interpreting and using explanations.

First, while participants who received explanations clearly outperformed those without explanations, participants who had access to both observations and explanations did not achieve higher accuracy than those with explanations alone. Our analysis suggests that people with observations and explanations often behave as though they relied exclusively on the explanations. One reason for this may be that participants were explicitly told in our experiment that explanations were guaranteed to be true, which may have encouraged overreliance on explanations (cf. Vasconcelos et al., 2023). It is also possible that people are naturally biased towards reasoning through explanations based on prior experiences.

Next, we found that participants less frequently changed correct inferences to incorrect ones when given explanations, suggesting that explanations may help participants keep from changing correct inferences to incorrect ones. Moreover, providing explanations on top of observations helped participants make correct inferences more often. Whereas it is difficult to be confident about connections based on observations alone, explanations provide greater certainty about which connections do or do not exist.

Lastly, participants interpreted more from explanations than is directly communicated (cf. Kirfel et al., 2022). They were generally more likely to infer that no connections existed than that the connections were reversed when given negative explanations. Moreover, there was an interaction that affected only participants in the 'observation + explanation' condition that slightly increased the proportion of responses that inferred reverse connections when the explanations were factual$^-$ and further decreased when the explanations were counterfactual$^-$. This suggests that people may be biased towards inferring causal connections when they can visually see co-occurrences and no connections when they see co-non-occurrences. People further extended their inferences from negative explanations to connections not mentioned in the explanations, such as inferring that C causes B after being told that A does not cause B.

While these results hint at interesting features of how humans interpret and integrate explanatory information, our experimental setup also has some limitations. First, explanations contain temporal information whereas observations do not, since only the resulting activations are presented. This raises the question of how much the difference resulted from the explanatory versus the temporal information. Second, even though explanations contain information about the activation states for two of the three components, it is not as visually salient as seeing the device light up. Similarly, the history table containing information from previous trials provided in the experiment displays the observational data graphically whereas the explanations require reading each row. These may have affected how effectively participants were able to integrate explanatory information across trials, a limitation that does not apply for the normative model.

We plan to address these limitations in future experiments by adding temporal information to the observations, and by including the activation information in the history table for all conditions. Also, investigating how people provide and interpret explanations by asking them to choose explanations themselves may offer insight into the pragmatics of explanation generation and understanding. Lastly, we have only looked at how people use explanations with observations and by themselves, but not with interventions, and we hope to integrate all three tools of causal reasoning in future studies.

## Acknowledgments

AN was supported by the NSF Graduate Research Fellowships Program. TG was supported by a research grant from the Stanford Institute for Human-Centered Artificial Intelligence (HAI).

## References

Bareinboim, E., Correa, J. D., Ibeling, D., & Icard, T. (2022). On Pearl's hierarchy and the foundations of causal inference. In *Probabilistic and causal inference: The works of Judea Pearl* (pp. 507–556).

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., Gerstenberg, T., Mayrhofer, R., & Lagnado, D. A. (2018). Time in causal structure learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(12), 1880–1910.

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), 708–731.

Bürkner, P.-C. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, *80*, 1–28.

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

Gerstenberg, T. (2022). What would have happened? counterfactuals, hypotheticals and causal judgements. *Philosophical Transactions of the Royal Society B*, *377*(1866), 20210339.

Gong, T., Gerstenberg, T., Mayrhofer, R., & Bramley, N. R. (2023). Active causal structure learning in continuous time. *Cognitive Psychology*, *140*, 101542.

Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. Part I: Causes. *The British Journal for the Philosophy of Science*, *56*(4), 843–887.

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, *107*(1), 65–81.

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, *57*(1), 227–254.

Kıcıman, E., Ness, R., Sharma, A., & Tan, C. (2023). Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*.

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

Lombrozo, T., & Vasilyeva, N. (2017). Causal explanation. *Oxford handbook of causal reasoning*, 415–432.

Meder, B., Gerstenberg, T., Hagmayer, Y., & Waldmann, M. R. (2010). Observing and intervening: Rational and heuristic models of causal decision making. *Open Psychology Journal*, *3*, 119–135.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, England: Cambridge University Press.

Rothe, A., Deverett, B., Mayrhofer, R., & Kemp, C. (2018). Successful structure learning from observational data. *Cognition*, *179*, 266–297.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*(3), 453–489.

Ullman, T. D., Stuhlmüller, A., Goodman, N. D., & Tenenbaum, J. B. (2018). Learning physical parameters from dynamic scenes. *Cognitive Psychology*, *104*, 57–82.

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023, apr). Explanations can reduce overreliance on ai systems during decision-making. *Proc. ACM Hum.-Comput. Interact.*, *7*(CSCW1).

Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 216–227.