UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Do as I explain: Explanations communicate optimal interventions

Permalink

https://escholarship.org/uc/item/7w09v5vk

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

Authors

Kirfel, Lara
Harding, Jacqueline
Shin, Jeong Yeon
et al.

Publication Date

2024

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Do as I explain: Explanations communicate optimal interventions

Lara Kirfel (kirfel@mpib-berlin.mpg.de), Center for Humans and Machines, Max Planck Institute for Human Development
Jacqueline Harding (hardingj@stanford.edu), Department of Philosophy, Stanford University
Jeong Shin (jyshin@stanford.edu), Department of Computer Science, Stanford University
Cindy Xin (cindywxx@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

People often select only a few events when explaining what happened. What drives people's explanation selection? Prior research argued that people's explanation choices are affected by event normality and causal structure. Here, we propose a new model of these existing findings and test its predictions in a novel experiment. The model predicts that speakers value accuracy and relevance. They choose explanations that are true, and that communicate useful information to the listener. We test the model's predictions empirically by manipulating what goals a listener has and what actions they can take. Across twelve experimental conditions, we find that our model accurately predicts that people like to choose explanations that communicate optimal interventions.

Keywords: explanations; causality; optimal intervention; pragmatic reasoning; normality.

Introduction

Imagine the following scenario: Your car's engine frequently fails to start. You need to keep it intact a bit longer, but you're hesitant to invest a lot of money into repairs since you plan on selling it soon. A friend of yours who's knowledgeable about cars inspects your car and identifies two potential causes for engine failure: worn out spark plugs and a damaged timing belt. Each issue individually can cause start-up failures, with repairs for each estimated at around $300. However, your friend also tells you that both factors differ in how often they actually cause an engine failure. While worn out spark plugs frequently cause engine failures, a damaged timing belt rarely does so. With this in mind and aiming for minimal repair, you take the car to the mechanic. After telling the mechanic that the car has been inspected by a friend of yours already and only needs repair, they ask you "Why did the engine fail?". How will you respond?

Speak Truthfully, Act Optimally

In our daily interactions, we constantly exchange explanations, whether clarifying a missed appointment, detailing how a gadget works, or sharing stories (Hilton, 1990). These exchanges are mostly seamless and quick, and adhere to systematic patterns. We strive for *truthfulness*, ensuring our explanations are based on accurate information. We aim for *relevance*, providing details pertinent to the topic of conversation. We offer just enough detail to be understood without overwhelming, and we strive for clarity. Often subsumed under the term of Gricean "Cooperative Principles" (Grice, 1989), these implicit guidelines help characterize how people communicate effectively.

When choosing how to best communicate, people often need to make "trade-offs" between these communicative principles (Sumers, Ho, Griffiths, & Hawkins, 2023). Speakers must balance being truthful and relevant depending on the listener's time, resources, and goals. Consider again the scenario from the start. Here, the speaker faces two distinct communicative pressures: Adhering to the maxim of truthfulness would mean informing the mechanic about *both* potential causes of the problem. This approach aligns with the principle of providing complete and accurate information, regardless of the immediate practical implications. Yet, at the same time it is in the speaker's interest to effectively guide the listener's action towards a cost-effective solution by mentioning only that cause in the explanation that frequently causes the issue. Given these two competing factors, what should a speaker say?

Be Relevant, But How?

While being truthful is a primary communicative objective (Moeschler, 2021), not everything that is true is also of interest for the listener. Grice's maxim of Relevance encourages speakers to say what's related to the "question under discussion" (Benz & Jasinskaja, 2017; Grice, 1957, 1975; Roberts, 2012). Several accounts of communication recognize language as a form of action, with an emphasis on the speaker's intentions and their effects on listeners (Austin, 1962, 1975; Frank & Goodman, 2012; Goodman & Frank, 2016; Searle, 2014). Rather than achieving a merely *cognitive* effect, the relevance of an utterance is measured by how it influences the listener's actions. In addition to truthfulness as an epistemic utility (Bridgers, Jara-Ettinger, & Gweon, 2020; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016), a speaker also considers an utterance's *decision-theoretic* utility (Benz, 2011; Benz & Van Rooij, 2007; Sumers et al., 2023). Relevant speech equips the listener with knowledge that allows them to act efficiently towards their goals (Hawkins, Stuhlmüller, Degen, & Goodman, 2015).

Selection of Causal Explanations

When explaining what happened and why, people often mention only a few causal factors (Hart & Honoré, 1959/1985; Tversky & Kahneman, 1973). People's explanation choices

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.* ©2024 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

are affected by event normality and the causal structure of the situation (e.g. Gerstenberg & Icard, 2020; Icard, Kominsky, & Knobe, 2017; Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015). They prefer to select an abnormal cause in conjunctive causal structures where each cause is necessary for the outcome (Güver & Kneer, 2023; Henne, O'Neill, Bello, Khemlani, & De Brigard, 2021; Kirfel & Lagnado, 2018; Kominsky & Phillips, 2019; Willemsen & Kirfel, 2019), but a normal cause in disjunctive causal structures where a single cause is sufficient (Gerstenberg & Icard, 2020; Icard et al., 2017). What explains this intriguing pattern?

Some argue that people prefer causal factors that reliably bring about outcomes, and are robust to possible changes in the background conditions (Grinfeld, Lagnado, Gerstenberg, Woodward, & Usher, 2020; Lombrozo, 2010; Morris et al., 2018; Quillien & Lucas, 2023; Vasilyeva, Blanchard, & Lombrozo, 2018; Woodward, 2006). Communicating such robust causes could be helpful as they suggest optimal points of intervention (Hitchcock, 2012). Morris et al. (2018) suggest that causal judgements accumulate knowledge about intervention effectiveness: by repeatedly judging whether an event was a cause of an outcome, people estimate the average likelihood that intervening on this cause would bring about the outcome. Take again the scenario from the beginning: In order to prevent the outcome (the engine failure) in a disjunctive causal structure from occurring, it is most effective to make sure the cause is disabled that is most likely (the "normal" causal) to bring about the outcome (faulty spark plugs). More generally, intervening on an abnormal cause in a conjunctive causal structure, and a normal cause in a disjunctive causal structure, makes the largest difference to the probability of the outcome (Kirfel, Icard, & Gerstenberg, 2022).

That said, identifying and communicating the best target of intervention is highly dependent on the listener's goals and knowledge (Kirfel et al., 2022). Previous experimental tasks investigating causal selection (e.g., billiard ball setups Gerstenberg & Icard, 2020 or vignettes with social agents Icard et al., 2017), underspecify the listener's goal (if there is a listener at all). A speaker who doesn't know whether the listener wants to generate or prevent an outcome, for example, would be uncertain about what to say. We argue that the observed interaction between normality and causal structure in people's causal judgments can be explained by assuming that speakers want to communicate decision-relevant information to others. To test this idea, we design an explanation task and systematically vary what the desired outcome is, and how it can be achieved optimally. This allows our optimal intervention account of explanation to make systematic predictions about people's explanation choices.

A Model of Explanation Choice

We assume that a speaker chooses explanations by trading off two goals: *accuracy* and *relevance*. Speakers want to give explanations that allow the listener to draw accurate inferences about what happened. Speakers also want to give explanations that are relevant for the action goals that the listener has. Specifically, they may choose to cite targets in their explanations that would be useful for the listener to intervene on.

To model these intuitions, we build on formal models of communication and apply them to explanations (Frank & Goodman, 2012; Goodman & Frank, 2016). In a recent generalization of the Rational Speech Act framework, Sumers et al. (2023) model speakers as selecting utterances not only by assessing their accuracy, but also for their downstream relevance to the listener. Formally,

$$P_{Speaker}(\mathsf{Utterance} \mid w) \propto \exp\Big[\beta \cdot (\lambda U_{Relevance}(\mathsf{Utterance} \mid w) + (1-\lambda) U_{Accuracy}(\mathsf{Utterance} \mid w))\Big].$$

where $P_{Speaker}(\mathsf{Utterance} \mid w)$ denotes the speaker's likelihood of producing a given Utterance in context w, $U_{Relevance}$ denotes the relevance of the response to the listener, $U_{Accuracy}$ denotes its accuracy. β and λ are free parameters in the model. β captures the speaker optimality – it determines how likely a speaker will choose the utterance with the highest expected value (from the set of possible utterances). λ captures the extent to which the speaker weighs relevance and accuracy when choosing their explanation. If λ is 0, then the speaker only cares about accuracy, and if λ is 1, then the speaker only cares about relevance. From this part of the model, we derive the following general hypothesis:

Hypothesis 1a): Speakers trade off accuracy and relevance when giving explanations.

Importantly, Sumers et al. (2023, p.14) propose that $U_{relevance}(\mathsf{Utterance})$ should be understood as the expected reward of the listener's "decision policy" after hearing Utterance. Formally,

$$U_{Relevance}(\mathsf{Utterance} \mid w) = \sum_a \pi_{Listener}(a \mid \mathsf{Utterance}) \cdot \mathsf{Reward}(a \mid w)$$

where $\pi_{Listener}(a \mid \mathsf{Utterance})$ denotes the listener's propensity to select an action a after hearing Utterance and $\mathsf{Reward}(a \mid w)$ denotes her reward from picking a in context w (the action space and rewards characterise the decision problem she faces). From this part of the model, we derive the following hypothesis:

Hypothesis 1b): People favor explanations that point to optimal interventions.

By treating the experimental setting of this paper as one such decision problem, we can use this model of communication to generate predictions for participant responses.

Experiment: Manipulating causal structure, outcome valence, and intervention type

Methods

Participants We recruited 590 participants (*age*: M = 36, SD = 13; *gender*: Female = 304, Male = 263, Non-binary =

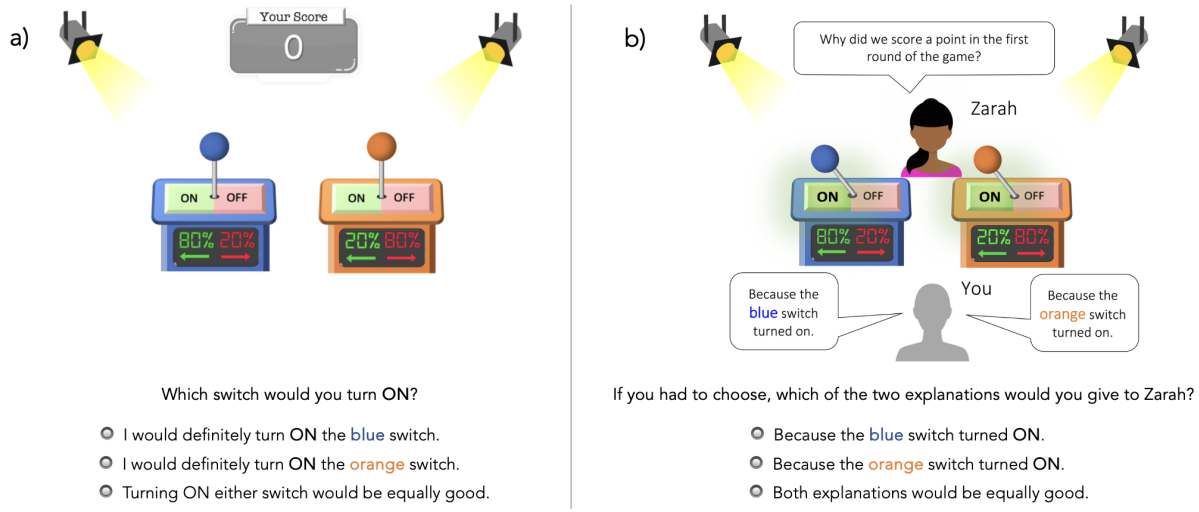

Figure 1: **Experimental Tasks** a) In the "Intervention Task", participants choose which switch they would like to intervene on to change it's probability of success. b) In the "Explanation task", participants choose what explanation to give to their teammate who gets to intervene on one of the switches in the next round. Importantly, the teammate doesn't know how likely each switch is to succeed by itself.

18; *race*: Asian = 50, Black/African American = 41, Multiracial = 41, White = 438, Other = 17), via Prolific (Palan & Schitter, 2018).

Design The experiment employs a 2 causal structure (conjunctive vs disjunctive, *within-subject*) × 2 outcome (positive vs negative, *between-subject*) × 3 intervention type (hard vs soft vs fixed, *between-subject*) design. Each participant is randomly assigned to one of six experimental conditions. Two of them include the hard intervention set up, two the soft intervention setup, and two the fixed intervention set up. For each intervention type, one condition contains a positive outcome, and one contains a negative outcome. Each condition contains both causal structures, in counterbalanced order. Study materials, design, and analyses were pre-registered: `https://github.com/cicl-stanford/explanation_intervention`.

Procedure In our experimental scenario, participants take part in the game show "Flip or Flop" (see Figure 1). In this game show, participants play in teams or two, and, depending on the game condition, the goal is either to avoid losing points (negative outcome condition) or to win points (positive outcome condition). Whether a point is lost or won is determined by the state of two switches. There is a blue switch and an orange switch, and both switches can either turn "ON" or "OFF". For example, in the conjunctive causal structure condition, it needs two switches to be turned "ON" in order to win a point (or lose a point). In the disjunctive causal structure condition, it is sufficient if only one switch is turned "ON" in order to win or lose a point. Whether a switch turns "ON" or "OFF" is determined probabilistically, and each switch has a different likelihood to turn on or off. The blue switch has an 80% chance to turn "ON" and a 20% chance to turn "OFF", and the orange switch has a 20% chance to turn "ON", and a 80% chance to turn "OFF". Each subject plays "Flip or Flop" together in a team with another player, Zarah (conjunctive structure structure), or Alice (disjunctive structure condition). The participant and the team player, e.g. Zarah, differ in terms of how much they know about the game. Crucially, while the outcome in the first game round is determined probabilistically, the team player has the chance to intervene on one of the switches in the second round. The team's goal is to score as many points as possible (or to avoid losing points). As an example of the experimental procedure, consider a participant in the positive outcome condition who sees the conjunctive structure first part.

Conjunctive Causal Structure *Introduction*. The subject participates in the experiment as one of the two players in the game show. As a game show participant, they have full knowledge about the game situation. That is, the participant knows that it takes both switches to be turned on in order to win a point. In addition, the participant is informed about the different likelihoods of the switches. Participants then will need to answer the first (of in total three sets of) comprehension check questions correctly in order to proceed in the experiment: Four questions on causal structure of the game and probabilities of the switches.

Intervention task. After answering these comprehension checks correctly, participants proceed to an intervention task. In this task, they are instructed that they are given one trial round in which they have the chance to intervene on one of the switches.

Types of Intervention In the *hard intervention* condition, participants can turn one of the switches ON (negative condi-

tion: OFF) manually. The other switch will turn ON or OFF based on its probability. Participants can select one out of three action options: i) "I would definitely turn ON the blue switch.", ii) "I would definitely turn ON the orange switch.", iii) "Turning ON either switch would be equally good." In the *soft intervention* condition, participants are instructed that they have the chance to press one of the buttons that are connected to the switches. By pressing a button, they can increase the probability of one of the switches turning ON by 20% (negative condition: decrease the probability by 20%). In the *fixed intervention* condition pressing a button means players can increase the probability of one of the switches turning ON to 90% (negative condition: reduce the probability to 10%), irrespective of its prior probability.

After having completed the intervention task, participants are told that their team player, Zarah, has only selective knowledge about the game set up. Zarah knows that it takes both switches to be "ON" in order to score a point. Zarah knows that both switches differ in how likely they turn "ON" and that pressing a button will change the probability of the switch that's connected to that button turning on to 100% (*soft*: increase its probability by 20% / *fixed*: change it to 90%). However, in contrast to the participant, Zarah does not know how likely the blue switch is to turn "ON" and how likely the orange switch is to turn "ON". Participants are then asked four questions about their own knowledge about the game setup, and four questions about their team player Zarah's knowledge about the game set up.

Explanation Task After this, participants proceed to the final scenario and explanation selection task of the game. In the first round of the game, the participant and Zarah observe what happens: both the blue and the orange switch turned on and one point is scored. Remember that Zarah doesn't know how likely each switch was to turn on. Just before Zarah is about to decide which button she should press in order to increase the probability of one switch, she is allowed to ask her teammate, i.e. the participant, for an explanation. She inquires about the round that just finished: "Why did we score a point in the last round?". Participants now have the task to choose one of two possible explanations, or indicate no explanatory preference, in a forced choice task: "We scored a point…" i) "... because the blue switch turned 'ON'", ii) "...because the orange switch turned 'ON'", iii) "both explanations are equally good".

Implementing Our Explanation Model

Earlier, we described a general model of communication from Sumers et al. (2023), and suggested it could be applied to our experimental setting. Here, we make good on this promise, concretizing the model and using it to generate predictions.

Corresponding to the key features of our experiments, let

$$\begin{aligned}\mathsf{Switch} &\in \{\mathsf{Blue}, \mathsf{Orange}\}\\ \mathsf{S} &\in \{\mathsf{Conjunctive}, \mathsf{Disjunctive}\}\\ \mathsf{C} &\in \{\mathsf{Positive}, \mathsf{Negative}\}\\ \mathsf{IT} &\in \{\mathsf{Hard}, \mathsf{Soft}, \mathsf{Fixed}\}.\end{aligned}$$

By $\mathsf{Reward}(\mathsf{Switch} \mid \mathsf{S}, \mathsf{C}, \mathsf{IT})$, we denote the expected reward to a participant who intervenes on Switch according to their team's condition (C), the intervention type (IT) and the causal structure (S). Recall that Blue (respectively, Orange) is the normal (respectively, abnormal) switch. For the case depicted in Figure 1, we have

$$\begin{aligned}&\mathsf{Reward}(\mathsf{Blue} \mid \mathsf{Conjunctive}, \mathsf{Positive}, \mathsf{Hard})\\ &\quad = 1 \cdot P(\text{Both switches are ON})\\ &\quad = P(\mathsf{Blue} = \mathsf{ON}) \cdot P(\mathsf{Orange} = \mathsf{ON})\\ &\quad = P(\mathsf{Orange} = \mathsf{ON})\\ &\quad = 0.2\end{aligned}$$

since a hard intervention on the blue switch for a team with the positive condition sets $\mathsf{Blue} = \mathsf{ON}$ with probability 1. The values of Reward across the different conditions are presented in Table 1.

We denote a participant's response

$$\mathsf{Response} \in \{\mathsf{Blue}, \mathsf{Orange}, \mathsf{No\ Preference}\}.$$

Note that the meaning of participants' responses changes across the intervention and explanation tasks. For the intervention task, $\mathsf{Response} = \mathsf{Blue}$ means that the participant indicates that they will themselves intervene on the blue switch in the next round. For the explanation task, $\mathsf{Response} = \mathsf{Blue}$ indicates that the participant provides her teammate the explanation "because the blue switch turned ON".

Table 1: Expected reward of taking different actions in the intervention task. Bold values show the unique best intervention in each situation. A = abnormal ($P(A) = 0.2$), N = normal ($P(N) = 0.8$). no pref = no preference (for this option, we assume a 50% chance of choosing to intervene in A or N). In our examples, we denoted N = blue switch, and A = orange switch (see Figure 1). For example, if the agent intervenes on the abnormal cause in the conjunctive, positive, hard condition, they have an 80% of scoring a point. Intervening on the normal cause only yields a 20% chance of success.

	conjunctive						disjunctive					
	positive outcome $P(A \wedge N) = P(A) \cdot P(N)$			negative outcome $1 - P(A \wedge N)$			positive outcome $P(A \vee N) = P(A) + Pr(N) - Pr(A \wedge N)$			negative outcome $1 - P(A \vee N)$		
	abnormal	*no pref.*	*normal*	*abnormal*	*no pref.*	*normal*	*abnormal*	*no pref.*	*normal*	*abnormal*	*no pref.*	*normal*
hard	**0.80**	0.50	0.20	1	1	1	1	1	1	0.2	0.5	**0.8**
soft	**0.32**	0.26	0.20	**1**	0.94	0.88	0.88	0.94	**1**	0.20	0.26	**0.32**
fixed	**0.72**	0.45	0.18	0.92	0.95	**0.98**	**0.98**	0.95	0.92	0.18	0.45	**0.72**

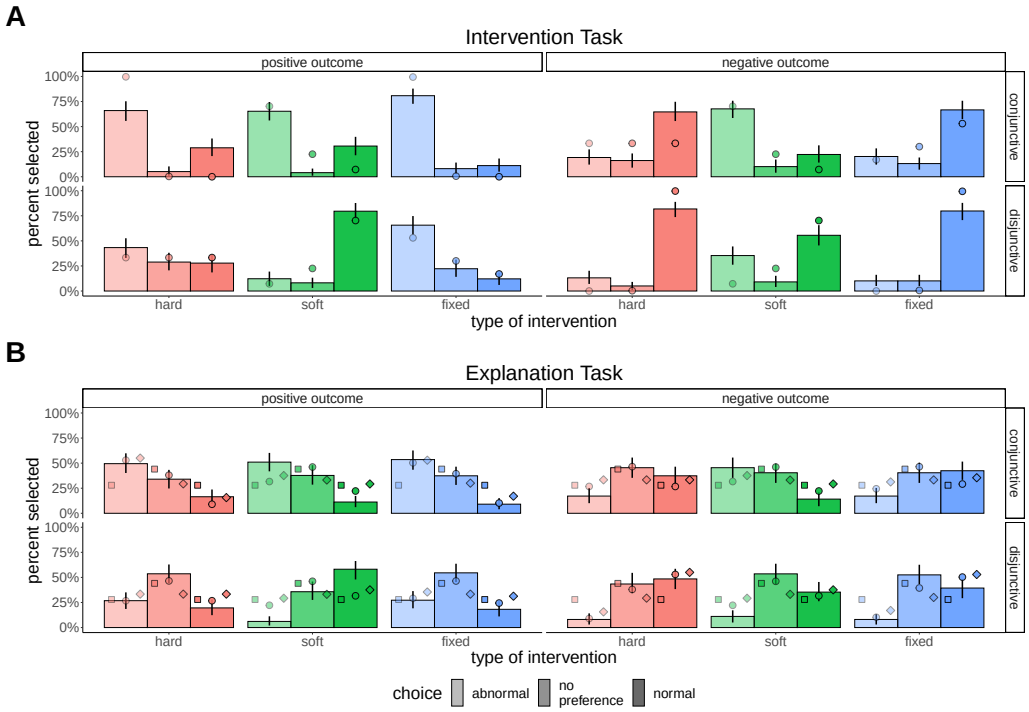

Figure 2: Participants' intervention choices (**A**) and explanation choices (**B**). The shading of each bar indicates the choice: light = abnormal, medium = no preference, dark = normal. Small shapes indicate model predictions ($\square$ = accuracy only model, $\circ$ = combined model, $\diamond$ = relevance only model). For example, the bar on the very left in the top panel shows the percentage of participants who selected the abnormal cause in the intervention task in the 'positive outcome' condition, with a 'conjunctive structure', when the type of intervention was 'hard'. *Note*: Error bars in all figures show bootstrapped 95% confidence intervals.

Intervention Task For the intervention task, we suggest that participants intervene on switches according to their expected reward (see Table 1). Specifically, for $\mathsf{Response} \in \{\mathsf{Blue}, \mathsf{Orange}\}$

$$P_{Intervention}(\mathsf{Response} = \mathsf{Switch} \mid \mathsf{S}, \mathsf{C}, \mathsf{IT}) \propto \exp(\beta \cdot \mathsf{Reward}(\mathsf{Switch} \mid \mathsf{S}, \mathsf{C}, \mathsf{IT}))$$

where $\beta \in [0, 50]$ is a (fitted) temperature parameter. We suppose that participants who say that either switch is equally good are indifferent between intervening on each, meaning

$$P_{Intervention}(\mathsf{Response} = \mathsf{No\,Preference} \mid \mathsf{S}, \mathsf{C}, \mathsf{IT}) \propto \exp\Big[\beta \cdot (0.5 \cdot \mathsf{Reward}(\mathsf{Blue} \mid \mathsf{S}, \mathsf{C}, \mathsf{IT}) + 0.5 \cdot \mathsf{Reward}(\mathsf{Orange} \mid \mathsf{S}, \mathsf{C}, \mathsf{IT}))\Big].$$

Explanation Task As discussed above, we follow Sumers et al. (2023) in supposing that participants will attempt to balance accuracy against relevance during communication:

$$P_{Explanation}(\mathsf{Response} \mid \mathsf{S}, \mathsf{C}, \mathsf{IT}) \propto \exp\Big[\beta \cdot (\lambda U_{Relevance}(\mathsf{Response} \mid \mathsf{S}, \mathsf{C}, \mathsf{IT}) + (1-\lambda) U_{Accuracy}(\mathsf{Response} \mid \mathsf{S}, \mathsf{C}, \mathsf{IT}))\Big].$$

Recall that $U_{Relevance}$ represents the listener's expected reward after hearing the speaker's utterance. To calculate $U_{Relevance}$ for $\mathsf{Response} \in \{\mathsf{Blue}, \mathsf{Orange}\}$, then, we simply suppose that the participant imagines their teammate intervening on the switch they name; the explanation's relevance is then given by the teammate's expected reward when making this intervention. So we have

$$U_{Relevance}(\mathsf{Switch} \mid \mathsf{S}, \mathsf{C}, \mathsf{IT}) \propto \exp(\beta \cdot \mathsf{Reward}(\mathsf{Switch} \mid \mathsf{S}, \mathsf{C}, \mathsf{IT})).$$

As with the intervention experiment, we assume that a response of "either explanation is equally good" indicates indifference between the two interventions their teammate could make (so $U_{Relevance}$ for this response is given by the average $U_{Relevance}$ of the other two choices). We define $U_{Accuracy}$ as:

$$U_{Accuracy}(\mathsf{Response} \mid \mathsf{S}, \mathsf{C}, \mathsf{IT}) = \begin{cases} 0 \text{ if } \mathsf{Response} \in \{\mathsf{Blue}, \mathsf{Orange}\} \\ 1 \text{ if } \mathsf{Response} = \mathsf{No\,Preference} \end{cases}.$$

We define $U_{Accuracy}$ in this way to reflect the fact that in all the cases we consider, the outcome's dependence on the switch value does not vary between switches (since both switches are on). So a speaker who wants to answer the why question as accurately as possible will not single out one switch over the other as *the* cause; both are causes.

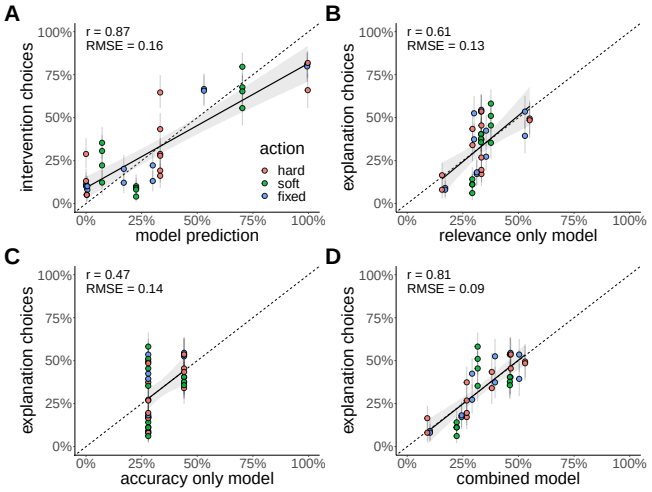

Figure 3: Scatter plots of model predictions and participant choices. **A** shows *intervention choices* (see Figure 2a). **B** to **D** show *explanation choices* with the predictions of different models (see Figure 2b).

We fit the speaker optimality parameter $\beta \in [0, 50]$, and the weighting parameter $\lambda \in [0, 1]$, obtaining $\beta = 3.51$ and $\lambda = 0.84$. Applying these values to the case depicted in Figure 1 (with a conjunctive structure, hard intervention and positive condition), we predict that the probability a participant selects the different options are: $p(\text{blue}) = 0.09$, $p(\text{orange}) = 0.53$, and $p(\text{no preference}) = 0.38$.

Results

We discuss the results of the intervention task and explanation task in turn (see Figure 1).

Intervention Task Figure 2a shows participants' intervention choices together with the model predictions. Overall, participants' choices are highly correlated with the optimal intervention model (see also Figure 3a). For example, in a conjunctive causal structure in which the outcome is positive and the team player can turn one switch on ("hard intervention") (top row, leftmost panel), the majority of participants chose to intervene on the abnormal cause (however, there are also some participants who choose to intervene in the normal cause in that scenario). When the structure is disjunctive, the outcome positive, and the intervention soft (increasing the probability of one cause by 20%), participants choose to intervene in the normal cause. In sum, there is no general preference for participants to intervene on a normal, or abnormal cause. It depends on all the factors we manipulated: outcome valence, causal structure, and type of intervention.

Explanation Task Figure 2b shows participants' explanation choices together with the predictions of three models. The *accuracy only model* assumes that people only care about communicating explanations that are accurate. This model drops the relevance term from the speaker utility function. The *relevance only model* assumes that people only care about communicating what's relevant, dropping the accuracy term from the utility function. Finally, the *combined model* incorporates both aspects. As Figure 3c–d show, the combined model performs markedly better than any of the lesioned models, suggesting that both components are critical for capturing speaker's explanation choices.

To illustrate, compare people's explanation choices with their intervention choices in the "conjunctive causal structure / positive outcome / hard intervention" condition. Even though intervening on the abnormal causal is optimal here, more people choose 'no preference' when giving an explanation. This boost in 'no preference' selections can be seen across all conditions. As another example, consider the "disjunctive causal structure / negative outcome / fixed intervention" condition. Here, participants intervene in the normal cause (as predicted by the optimal intervention model) but, when giving explanations, they are most likely to select 'no preference'.

General Discussion

People choose explanations that communicate optimal interventions. Rather than generally citing abnormal causes in conjunctive structures and normal causes in disjunctive ones, our participants communicated the causal factor that was most optimal to intervene on when giving an explanation to their teammate. This includes not indicating a preference to cite one cause over the other when both causes are equally optimal to intervene on. However, rather than strictly citing the causal factor that is most optimal to intervene on, people trade-off relevance and accuracy in their explanations. People were generally more likely to express 'no preference' when choosing explanations than when choosing interventions. When choosing what explanation to give, speakers value accuracy (i.e., giving an explanation that's true) and relevance (i.e., giving an explanation that's useful; see Sumers et al., 2023). In our experimental setup, it was clear to the speaker what the listener's goals and possible actions were. In everyday situations, however, a speaker is likely to be unsure about either of both of these. According to our model, this uncertainty would be reflected in the explanations they choose. Our work contributes to the debate about what "relevance" means in the context of communicating explanations. In our model, "relevant" simply means pointing toward useful actions. These actions might take the simple form of throwing a switch in a game show, but they could also be more complex, such as blaming or punishing another person for what they did (or failed to do; see Alicke, Rose, & Bloom, 2012; Samland & Waldmann, 2016; Sarin & Cushman, 2024; Sytsma, 2020). From our model's perspective, people blame some but not others because they believe that it's the blameworthy person one should do something about – their behavior needs to change. Event normality, causal structure, and responsibility have all been argued to affect what explanations people choose. We suggest a simple unification: people care about optimal interventions when communicating explanations.

Acknowledgments

Our work was supported by a grant from the Stanford Institute for Human-Centered Artificial Intelligence (HAI).

References

Alicke, M. D., Rose, D., & Bloom, D. (2012). Causation, norm violation, and culpable control. *The Journal of Philosophy*, *108*(12), 670–696.

Austin, J. (1962). *Speech acts.* Oxford.

Austin, J. L. (1975). *How to do things with words* (Vol. 88). Oxford university press.

Benz, A. (2011). How to set up normal optimal answer models. *Language, games, and evolution*, *6207*, 14–39.

Benz, A., & Jasinskaja, K. (2017). *Questions under discussion: From sentence to discourse* (Vol. 54) (No. 3). Taylor & Francis.

Benz, A., & Van Rooij, R. (2007). Optimal assertions, and what they implicate. a uniform game theoretic approach. *Topoi*, *26*, 63–78.

Bridgers, S., Jara-Ettinger, J., & Gweon, H. (2020). Young children consider the expected utility of others' learning to decide what to teach. *Nature human behaviour*, *4*(2), 144–152.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

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

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, *20*(11), 818–829.

Grice, H. P. (1957). Meaning. *The philosophical review*, *66*(3), 377–388.

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Grice, P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.

Grinfeld, G., Lagnado, D., Gerstenberg, T., Woodward, J. F., & Usher, M. (2020). Causal responsibility and robust causation. *Frontiers in Psychology*, *11*, 1069.

Güver, L., & Kneer, M. (2023). Causation, foreseeability, and norms. In *Proceedings of the 45th annual meeting of the cognitive science society* (pp. 888–895).

Hart, H. L. A., & Honoré, T. (1959/1985). *Causation in the law*. New York: Oxford University Press.

Hawkins, R. X., Stuhlmüller, A., Degen, J., & Goodman, N. D. (2015). Why do you ask? good questions provoke informative answers. In *Cogsci.*

Henne, P., O'Neill, K., Bello, P., Khemlani, S., & De Brigard, F. (2021). Norms affect prospective causal judgments. *Cognitive Science*, *45*(1), e12931.

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

Hitchcock, C. (2012). Portable causal dependence: A tale of consilience. *Philosophy of Science*, *79*(5), 942–951.

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

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences*, *20*(8), 589–604.

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

Kirfel, L., & Lagnado, D. A. (2018). Statistical norm effects in causal cognition. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 615–620). Austin, TX: Cognitive Science Society.

Kominsky, J. F., & Phillips, J. (2019). Immoral professors and malfunctioning tools: Counterfactual relevance accounts explain the effect of norm violations on causal selection. *Cognitive science*, *43*(11), e12792.

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

Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive psychology*, *61*(4), 303–332.

Moeschler, J. (2021). Why truth matters: When relevance meets truthfulness. *Pragmatics & Cognition*, *28*(2), 416–440.

Morris, A., Phillips, J., Icard, T., Knobe, J., Gerstenberg, T., & Cushman, F. (2018). Judgments of actual causation approximate the effectiveness of interventions. *Psy ArXiv) doi*, *10*.

Palan, S., & Schitter, C. (2018). Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

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

Roberts, C. (2012). Information structure: Towards an integrated formal theory of pragmatics. *Semantics and pragmatics*, *5*, 6–1.

Samland, J., & Waldmann, M. R. (2016). How prescriptive norms influence causal inferences. *Cognition*, *156*, 164–176.

Sarin, A., & Cushman, F. (2024). One thought too few: An adaptive rationale for punishing negligence. *Psychological Review*, *131*(3), 812.

Searle, J. (2014). What is a speech act? In *philosophy in america* (pp. 221–239). Routledge.

Sumers, T. R., Ho, M. K., Griffiths, T. L., & Hawkins, R. D. (2023). Reconciling truthfulness and relevance as epistemic and decision-theoretic utility. *Psychological Review*.

Sytsma, J. (2020). Causation, responsibility, and typicality. *Review of Philosophy and Psychology*.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, *5*(2), 207–232.

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

Willemsen, P., & Kirfel, L. (2019). Recent empirical work on the relationship between causal judgements and norms. *Philosophy Compass*, *14*(1), e12562.

Woodward, J. (2006). Sensitive and insensitive causation. *The Philosophical Review*, *115*(1), 1–50.