UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Causal language about social interactions

Permalink
https://escholarship.org/uc/item/2xj92240

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors
Teo, Verona
Bergey, Claire Augusta
Gerstenberg, Tobias

Publication Date
2026

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

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

Causal language about social interactions

Verona Teo[1*], Claire Augusta Bergey[2] & Tobias Gerstenberg[1]
[1]Department of Psychology, Stanford University
[2]Department of Linguistics, Stanford University
*veronateo@stanford.edu

Abstract

Causal language is central to our understanding of social interactions—whether someone "caused" or "allowed" another's action shifts our impression of what happened. Yet models of causal language use have largely focused on physical events (e.g., billiard balls), ignoring the beliefs and preferences implicated in human action. We present a computational framework and three experiments investigating how people use causal expressions ("caused," "enabled," "allowed," "made no difference") across physical, epistemic, and preference-based interventions between agents. We find that people prefer different causal expressions across these intervention types: they describe removing a physical obstacle as a different form of facilitation than providing information. We capture people's language use with a model that selects utterances based on counterfactual simulations of events, inferences about agents' mental states, and utterance informativity. This model explains human judgments better than baseline models, suggesting that describing social influence involves reasoning about mental states, alternative actions, and alternative utterances.

Keywords: causal language; semantics; pragmatics; counterfactual simulation; social cognition

Introduction

Causal language shapes how we attribute responsibility and assign blame. Consider the distinction between saying that an action "caused" harm versus saying that it "allowed" it. Although both utterances may describe the same scenario, they highlight different aspects of the causal structure, evoking different mental representations of the event and how it came about. Philosophers (Foot, 1967, 2002; McDermott, 1995; McGrath, 2003; McMahan, 1993; Thomson, 1976), linguists (Baglini & Siegal, 2021; Nadathur & Lauer, 2020; Talmy, 1988; Wolff, 2003), and psychologists (Goldvarg & Johnson-Laird, 2001; Sloman et al., 2009; Wolff, 2007) have been interested in how people use and understand causal expressions.

To explain how people choose among these expressions, pragmatic theories of language use posit that language users consider not just semantics, the literal meaning of words, but also pragmatics, inferences about an utterance's informativity in context (Grice, 1975). That is, speakers prefer not just true but also contextually informative expressions. For example, when ball A collides with ball B and B ricochets, it is both true that A "affected" B's motion but also that it "caused" B's motion. Given that "caused" is more specific than "affected," a speaker will likely prefer to say the more informative "caused." Recent work has applied this theory to causal language, enumerating semantic meanings of causal terms and modeling people as pragmatically reasoning over them. Beller and Gerstenberg (2025) developed a model in the Rational Speech Acts (RSA) framework (Degen, 2023; Frank & Goodman, 2012) that casts speakers as considering alternative potential event outcomes (counterfactual simulations of what happened) and alternative potential utterances (pragmatic reasoning about informativity). They found that people choose the causal expressions that are most informative to a listener against the backdrop of relevant alternatives. This work focused on physical causation, such as scenarios involving billiard balls and direct physical contact.

In addition to this work in the physical domain, others have also attempted to formalize causal language in social contexts. Cao et al. (2023) developed a formal semantics for causal verbs using structural causal models, showing how expressions like "caused," "enabled," and "prevented" map onto distinct patterns of counterfactual dependence. However, this work also focused primarily on situations where an agent intervenes directly on the physical environment (e.g., one agent places or removes obstacles that affect what another agent is able to do). Further, it aimed to capture the semantics of different expressions (whether people think an expression is true), and not their pragmatics (which expression people think best describes what happened in context).

Causal language about social interactions needs to consider the agents' underlying mental states. For instance, one person can influence another not just by altering the physical environment, but also by changing their beliefs (e.g., by providing important information) or preferences (e.g., by making some options more or less attractive). Such influence can be exercised intentionally, raising questions about how causal language reflects not only what physically happened, but also the intervener's knowledge and goals. How do people integrate these factors when choosing how to describe what happened?

In this paper, we investigate how people use causal language across three types of agent interventions: physical (affecting what outcomes are physically accessible), epistemic (affecting what agents believe), and preference (affecting what agents prefer by changing option utilities). We focus on four causal expressions that span different types of causal involvement: "caused," "enabled," "allowed," and "made no difference." We develop a computational model that extends prior work to integrate counterfactual simulation, mental state inference, and pragmatic reasoning to predict how people select among these

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* ©2026 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

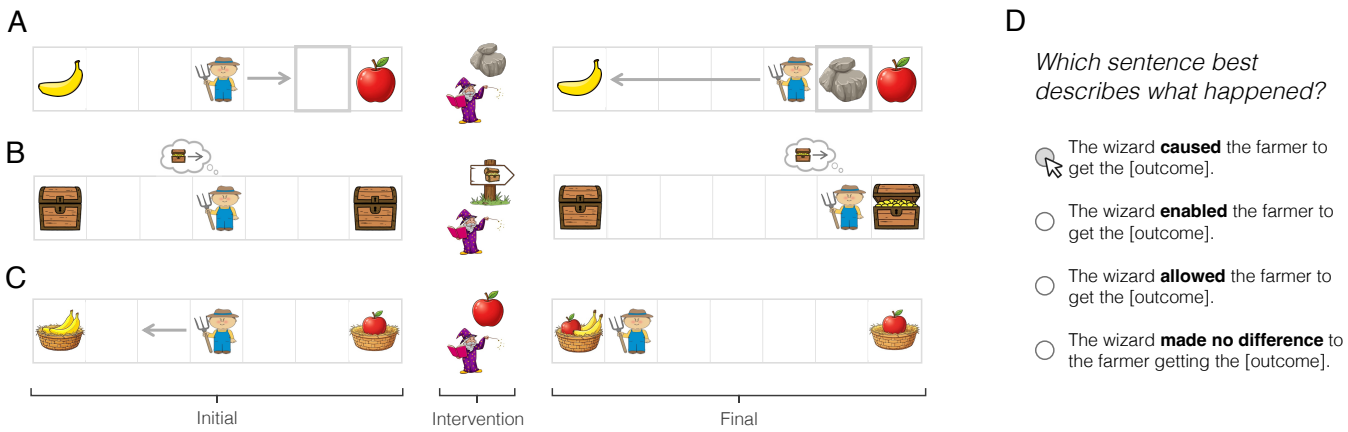

Figure 1: **Overview of the three experiments**. Each scenario involves the farmer taking an initial action or having an initial belief, the wizard's choice of intervention, and a final outcome. **(A)** In an example scenario from the *physical* domain, the farmer initially goes toward the apple before the wizard decides to block it with the rock. The farmer is then redirected and gets the banana. **(B)** In the *belief* domain, the farmer has an initial guess that the gold is on the right; the wizard reinforces that belief by showing a signpost indicating that the gold is to the right; the farmer continues going to the right and gets the gold chest. **(C)** In the *preference* domain, the farmer initially goes toward the left basket with two bananas. The wizard adds an apple to the left basket, and the farmer decides to get the left basket. **(D)** In each experiment, participants choose one of four utterances (where each utterance contains one of "caused," "enabled," "allowed," or "made no difference") that best describes the given scenario.

expressions when describing social interactions across the three intervention mechanisms. We find that causal language use in social contexts depends not only on changes to the physical causal structure, but also on inferred beliefs and preferences.[1]

Experimental Design

We conducted three experiments investigating how people choose causal expressions to describe scenarios involving a farmer and a wizard (Figure 1). The farmer seeks to obtain some outcome, and the wizard has the ability to influence the ultimate outcome by intervening in the environment. Across all experiments, participants watched animated scenarios and then selected which of four causal expressions best described what happened.

We focus on three distinct types of interventions that span the principal ways an agent can causally influence another's outcomes in social settings. First, in the *physical* domain (Experiment 1), the wizard directly manipulates the environment by adding or removing an obstacle that affects what the farmer can physically access. This provides a baseline for understanding causal language in settings similar to those in prior work examining physical interventions (Cao et al., 2023). Second, in the *belief* domain (Experiment 2), the wizard intervenes on the farmer's epistemic state by providing information about the world. This allows us to examine whether causal expressions track outcomes differently when causation operates through belief change rather than physical affordances, and introduces questions about knowledge asymmetry and trust. Third, in the *preference* domain (Experiment 3), the wizard influences the farmer's choices by changing the relative desirability of available options. Here, the farmer maintains both physical access and accurate knowledge, but the wizard alters which option is more attractive. Together, these three intervention types allow us to investigate how the same causal expressions—"caused," "enabled," "allowed," and "made no difference"—map onto different mechanisms of social influence.

[1]Code and materials are available here: `https://github.com/veronateo/social-causal-lang`

Computational Model

We propose a model that aims to capture causal language use via the Rational Speech Acts (RSA) framework. We first discuss the model's semantics, then discuss the pragmatic reasoning process over causal expressions.

World Model

We hypothesize that the meaning of each causal expression depends on three key features: counterfactual dependence (C), action (A), and value alignment (V). We represent world states w as a tuple of these features, where $C \in [0, 1]$ represents the degree to which the outcome counterfactually depended on the wizard's action, $A \in \{0, 1\}$ indicates whether the wizard actively intervened (1) or did nothing (0), and $V \in [0, 1]$ represents the degree to which the final outcome aligns with what the farmer values.

Mental State Inference We model the farmer as a rational agent planning under uncertainty about potential wizard intervention. The farmer plans actions a to maximize expected utility:

$$U(a; \theta) = \mathbb{E}_{o \sim P(o|a)} \left[R(o; \theta) - \kappa(o, a) \right] \quad (1)$$

where $P(o|a)$ represents the farmer's belief about which outcome o will result from action a, $R(o; \theta)$ assigns reward

$\theta \in [0, 1]$ to one outcome and $1 - \theta$ to the alternative, and $\kappa(o, a) = c_{\text{step}} \cdot |\text{path}(a, o)|$ is the path cost, which scales with the number of steps required to reach the outcome. We assume that the farmer has uniform beliefs about wizard intervention (i.e., the probability that the wizard acts is 0.5). The farmer chooses actions according to a softmax distribution with temperature parameter τ:

$$P(a \mid \theta) \propto \exp\left(\frac{U(a; \theta)}{\tau}\right) \tag{2}$$

where τ is a temperature parameter that controls how strictly the farmer selects actions that maximize expected utility.

We infer the farmer's preference θ by computing the posterior $P(\theta \mid a_{\text{obs}})$ given observed actions a_{obs} via Bayesian inference. We then compute the value alignment (V) as the probability that the actual outcome corresponds to the farmer's preferred outcome, marginalizing over the posterior distribution of θ.

Causal Inference We compute the counterfactual dependence (C) of the outcome on the wizard's (in)action (i.e., the causal relevance of the wizard's decision) via counterfactual simulation. Specifically, we compare the actual observed world to the set of alternative worlds in which the wizard acted differently. For each alternative action, the model simulates the counterfactual outcome, accounting for the farmer's behavior by marginalizing over the posterior distribution of inferred mental states. We define C as the probability that the outcome would have been different, averaged across the alternative actions available to the wizard. This allows us to capture the causal status of an action relative to the set of possible interventions.

Semantics

We define the semantic meaning $[\![u]\!]$ of each utterance u in a given world state $w = (C, A, V)$ as a continuous value in $[0, 1]$, as follows.

"Caused" The wizard *caused* the outcome to occur if it counterfactually depended on the wizard's action and the wizard actively intervened.

$$[\![\text{“caused”}]\!] = C \cdot A \tag{3}$$

"Enabled" The wizard *enabled* the outcome if it counterfactually depended on his action, the wizard actively intervened, and the outcome was aligned with the farmer's preferences.

$$[\![\text{“enabled”}]\!] = C \cdot A \cdot V \tag{4}$$

"Allowed" The wizard *allowed* the outcome to occur if it counterfactually depended on his action (or inaction) and the outcome was aligned with the farmer's preferences.

$$[\![\text{“allowed”}]\!] = C \cdot V \tag{5}$$

"Made no difference" Finally, the wizard *made no difference* to the outcome if it did not depend on the wizard's action, that is, the wizard's action was not causally relevant.

$$[\![\text{“made no difference”}]\!] = 1 - C \tag{6}$$

These values are normalized across utterances to form a probability distribution before being passed to the pragmatics component of the model.

Pragmatics

Given that multiple expressions may apply to a world state, we model speakers as choosing expressions that are not only literally true, but also informative in context. Specifically, we take the pragmatic speaker S_1 to be a rational agent who selects an utterance u to communicate the world state w to a literal listener L_0. The literal listener L_0 interprets utterances by computing a posterior over world states based on the semantic truth values of each utterance u:

$$L_0(w \mid u) \propto [\![u]\!] \cdot P(w) \tag{7}$$

where $P(w)$ is a uniform prior over world states. The expected utility of an utterance u is

$$U(u; w) = \log(L_0(w \mid u)) - \kappa(u) \tag{8}$$

We assume a uniform cost $\kappa(u) = 0$ for all utterances. The pragmatic speaker S_1 then selects utterances to maximize informativity:

$$S_1(u \mid w) \propto \exp(\alpha \cdot U(u; w)) \tag{9}$$

where $\alpha > 0$ is a rationality parameter. Higher values of α lead to more deterministic selection of the optimal utterance.

The model has three free parameters: the rationality parameter α in the pragmatics module (Equation 9), and two inference parameters, the farmer's step cost c_{step} (Equation 1) and softmax temperature τ (Equation 2). We fit the model parameters by minimizing the negative log-likelihood of the participants' utterance choices under the model predictions. We found optimal parameter values of $c_{\text{step}} = 0.949, \tau = 0.193, \alpha = 0.446$.

Alternative Models

To test which components of our model are necessary for capturing causal language use, we compared our full model against several ablated alternatives. Each alternative removes a component, allowing us to assess specific hypotheses about what drives people's choice of expressions. We evaluated each model's performance using 5-fold cross-validation.

No Causal Inference This "No CI" model removes counterfactual simulation, that is, the variable C from the semantics. It provides a test of whether counterfactual reasoning is a core semantic feature of causal language.

No Mental State Inference This "No MSI" model does not infer the farmer's beliefs or preferences, ignoring the variable V. This tests whether mental state reasoning like inferring the farmer's underlying preference is necessary for choosing causal expressions in such social contexts.

No Pragmatics To assess the contribution of pragmatic reasoning, we also compare against a "No Prag" model that uses only the literal semantics of the causal language. This model assumes a speaker who chooses utterances proportional to their literal truth values, without considering informativity: $S_0(u \mid w) \propto [\![u]\!]$. This tests whether pragmatic inference plays a role in causal language selection, or whether people simply choose expressions based on their literal meanings.

Figure 2: **Participant selections and model predictions across experiments**. Empirical distributions of human utterance choices and corresponding full model predictions for the aggregate distribution across all trials within each domain (above dashed line), followed by three example scenarios from each domain (below dashed line).

Methods

Participants

We recruited 100 participants for each experiment via Prolific, for a total of $N = 300$ participants (*age*: mean = 44, SD = 13; *gender*: 166 female, 131 male, 3 undisclosed; *race*: 222 white, 29 Black, 20 Asian, 16 mixed, 10 other). We included participants who were fluent in English, based in the United States, had an approval rate of at least 97%, and had completed at least 100 previous studies. Participants were compensated at a rate of $12 per hour.

Stimuli

All stimuli were animated scenarios showing a farmer and a wizard (Figure 1). In each scenario, the farmer started at the center of a horizontal lane, and could move left or right to reach one of two goal states. The wizard was positioned underneath the lane. Each trial began by revealing the farmer's initial state (e.g., steps toward one goal, beliefs about the options indicated by thought bubbles), followed by the wizard's intervention, and concluded with the farmer's final action and outcome. Across all three experiments, we did not include trials that would have suggested that the farmer was irrational (e.g., moving towards the apple when the rock is not initially present, then switching to go to the banana even if the wizard did nothing). This gave us 38 trials in total (*physical*: 10 trials; *belief*: 14 trials; *preference*: 14 trials).

Physical In this domain, the farmer navigated toward one of two fruit baskets (banana on the left, apple on the right; see Figure 1A). A bolded tile in front of the apple marked where the wizard could place or remove a rock. The farmer took initial steps toward one of the fruits. After observing the farmer's initial movement, the wizard then either placed a rock blocking the apple, removed an existing rock in front of the apple, or did nothing. The farmer continued moving, depending on which path was available, and ultimately reached either the apple or the banana. Across 10 trials, we systematically varied the farmer's initial direction, whether a rock was initially present, the wizard's action, and the final outcome.

Belief Here, closed treasure chests sat on both sides of the lane, with one containing gold and the other rocks (Figure 1B). The farmer could not see what was inside until he opened one of the chests. Thought bubbles displayed the farmer's initial belief about the gold's location; the farmer could either have no belief, a true belief, or a false belief. After the farmer's belief was displayed, the wizard could show a sign pointing correctly to the gold, a false sign pointing to the rocks, or do nothing. The farmer could then either follow the sign, ignore it, or (if the wizard didn't show a sign) make a guess, before ultimately obtaining either gold or rocks. We manipulated the farmer's initial belief state, the wizard's action, the farmer's subsequent response, and the outcome across 14 trials.

Preference In these scenarios, there were two fruit baskets on each side of the lane, with the left basket initially containing bananas and the right containing apples (Figure 1C). After the farmer took initial steps toward one basket, the wizard could add an apple to the left basket, add an apple to the right basket, or do nothing. The farmer then either continued toward his initial choice or switched baskets. The 14 trials varied the farmer's initial direction, the wizard's action, whether the farmer switched directions, and which basket the farmer ultimately obtained.

Procedure

All experiments followed the same general procedure. Participants first read instructions explaining the experimental setup and viewed demonstration animations showing each of the wizard's possible actions. Then, participants completed a comprehension check before proceeding to the main trials.

Results

Behavioral

Across the experiments, we observed distinct patterns of causal language use, finding that people's utterance choices are sensitive to the mechanism of intervention as well as the agents' mental states and causal roles (see Figure 2).

Physical In the physical intervention setting, participants overwhelmingly chose "caused" to describe scenarios in which the wizard prevented the farmer from reaching the apple. People strongly preferred to say that the wizard "caused" the farmer to get the banana when he was redirected due to the wizard either placing a rock or not removing the rock if it was initially present. In contrast, for cases in which the farmer achieved a seemingly positive outcome—outcomes that seemed to align with his preferences—participants were generally split between "enabled" and "allowed." People slightly preferred to say that the wizard "enabled" the farmer to get the apple when the wizard removed the rock and the farmer switched directions from the banana to the apple, and "allowed" specifically when the wizard chose not to block the farmer from reaching the apple.

Belief In the belief domain, people's choices depended on how the wizard's action influenced the farmer's epistemic state, and how the farmer responded to such information. When the wizard provided information and the farmer followed the advice, people most frequently chose "caused" and "enabled." Interestingly, unlike in the physical domain in which people mostly only use "caused" when the wizard's action worked against the farmer's goal, people also preferred choosing "caused" when the wizard's information led the farmer to successfully find the gold.

Furthermore, people chose "enabled" when the wizard reinforced the farmer's correct guess about the gold's location, as well as in situations where the wizard's action led the farmer to an undesirable outcome. This contrasts with the physical intervention case, in which people chose "enabled" only when the outcome was aligned with the farmer's preferences, potentially suggesting that in epistemic contexts, "enabled" tracks whether the agent was empowered to make an informed choice, regardless of whether the outcome was desirable.

Table 1: **Model performance comparison**. Negative log-likelihood (NLL) is evaluated on held-out test folds via 5-fold cross-validation (mean ± SE across folds split by trial). Jensen-Shannon Divergence (JSD) and root mean square error (RMSE) are computed on the full dataset after fitting the parameters to all trials (mean with bootstrapped 95% CIs across trials).

Model	NLL	JSD	RMSE
Full	**0.951 ± 0.048**	**0.073 [0.051, 0.095]**	**0.150 [0.116, 0.180]**
No CI	1.219 ± 0.020	0.179 [0.147, 0.213]	0.265 [0.238, 0.292]
No MSI	1.098 ± 0.063	0.140 [0.107, 0.177]	0.228 [0.195, 0.261]
No Prag	1.229 ± 0.124	0.102 [0.072, 0.138]	0.181 [0.144, 0.217]

Preference Finally, in the preference domain, people's utterance choices depended on whether the farmer's choice was changed due to the wizard's action. When the wizard's intervention led the farmer to switch options, participants strongly preferred "caused." However, when the wizard did nothing or added an apple to the option that the farmer was already walking toward, "made no difference" was the dominant response. In the cases in which the wizard "reinforced" the farmer's choice by adding an apple to the basket that the farmer was already going toward, people also sometimes chose "enabled."

Model Comparisons

Figure 2 shows the full model predictions compared to human judgments across all trials and for three select trials across the physical, belief, and preference experiments. Overall, the full model achieves the best performance across several metrics (Table 1, Figure 3). However, particularly in the belief experiment, the model often overpredicts "allowed" and underpredicts "made no difference." This is likely because the counterfactual dependence factor C averages over all alternative possible interventions. For scenarios in which the farmer has the correct initial belief about the gold's location and the wizard decides not to act, the dependence score is moderate rather than zero because the outcome might have differed if the wizard had shown the incorrect sign. The model interprets this moderate causal impact combined with the farmer achieving a positive outcome as "allowed," whereas humans may view the wizard's inaction as making no difference.

Compared to the full model, the "No Pragmatics" model performs worse, indicating that people selectively choose expressions that are more informative, even when several expressions are literally true. For instance, in the preference domain, an action that makes the farmer switch baskets might be literally compatible with "caused," "enabled," and "allowed." In these cases, the full model with pragmatics upweights "caused" because it is more informative, whereas the semantics-only model spreads the probability mass more evenly across all valid terms.

Similarly, the "No Mental State Inference" model also does not capture human judgments well, suggesting that inferring the farmer's values and deciding whether the wizard's intervention aligns with such beliefs and preferences contribute to better predictions of people's utterance choices. Because this model does not take into account whether the wizard's action was aligned with the farmer's preferences, it often predicted "caused" and "enabled" equally. However, people's utterance choices often distinguished between these terms, particularly in the physical domain.

Finally, the "No Causal Inference" model performs the worst, highlighting that counterfactual reasoning is a significant aspect of people's utterance choices. Without assessing how important the wizard's action was to the farmer's outcome, this model tends to over-attribute the wizard's causal role. Even in cases in which the wizard's action was irrelevant, the model often predicted "caused," "enabled," or "allowed."

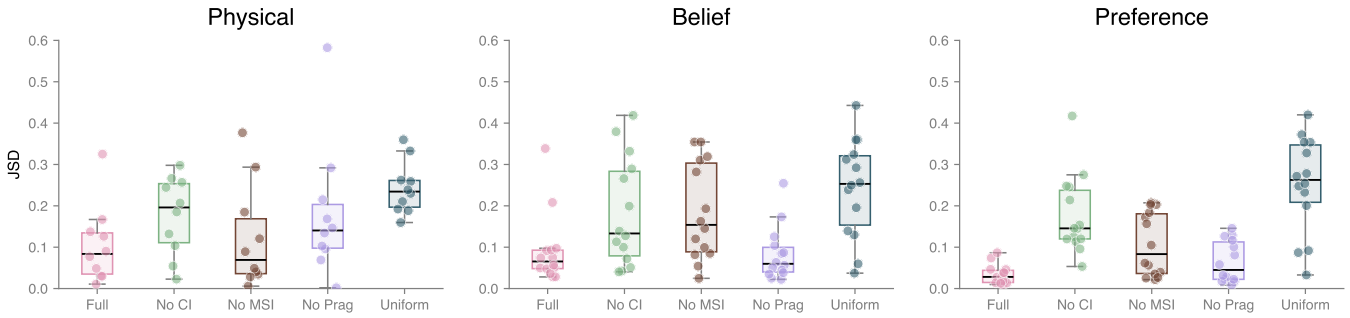

Figure 3: **Overall performance of each model**. Each point represents the Jensen-Shannon Divergence (JSD) between model predictions and human responses for a single trial (lower values are better). We also include a uniform baseline that assigns equal probability to each expression.

Discussion

In this work, we characterize how people use causal language to describe scenarios involving agents interacting with one another. We present a set of experiments that involve goal-directed agents in which one agent can influence another's outcome. We model people's behavior using a computational model of causal language use that unifies various social scenarios involving different intervention mechanisms (physical, epistemic, and preference). By integrating counterfactual reasoning with mental state inference and pragmatics, our model captures the graded and context-sensitive way in which people choose among a set of expressions ("caused," "enabled," "allowed," and "made no difference"). Our findings emphasize that people reason about agents' mental states, about what might have happened had the agent acted differently, and about what else a speaker could have said instead.

In extending causal language to social interaction settings, we find that people use causal expressions differently than in the physical scenarios previously studied. In purely physical settings, Beller and Gerstenberg (2025) posited that causing is a more specific instance of enabling—saying one ball's movement "caused" another ball's movement entails that it "enabled" it. Our findings suggest a different story in the social domain: "caused" acts as a broad term for causally relevant interventions, whereas "enabled" is more specifically used for interventions that successfully align with an agent's goal or preference. That is, among social agents, one cannot enable another to do what they did not want to do. People also reason about the set of alternative actions at an agent's disposal in each particular scenario. In the preference scenarios, in which the wizard had no way to block or deceive the farmer, participants rarely chose "allowed." In other words, if preventing another from getting their goal is not in the set of relevant actions, people do not find it appropriate to say that one agent "allowed" another to get to their goal by doing nothing. Social context, specifically the alignment of outcomes with agents' desires, shapes the meaning of causal language.

Complementing prior work, our results also suggest that the meanings of causal expressions are pragmatically adjusted based on the valence of the outcome. For instance, in the physical intervention experiment, participants tended to use "caused" for outcomes with negative utility (such as placing an obstacle so that the farmer got a less preferred fruit), but not positive utility (such as removing a rock so that the farmer got a more preferred fruit). In the belief experiment, people also sometimes used "caused" when the wizard's information changed the farmer's outcome positively (such as when he correctly informed the farmer and the farmer found the gold). That is, people found it appropriate to say that the wizard "caused" a positive outcome when he removed an informational obstacle, but not when he removed a physical obstacle.

Beyond variation across different types of interventions, individual participants also had differing intuitions; there was not a clear majority-choice causal expression for many scenarios. In real-world scenarios, speakers may choose among true expressions based on how they prefer to frame the event (e.g., exaggerate or minimize it, emphasize an agent's competence, or downplay an agent's role; see Macuch Silva et al., 2024; Rogers et al., 2017), and different speakers may resolve this lexical uncertainty differently (Bergen et al., 2016). In our task, with no particular reason to blame or praise the wizard or farmer, participants found several descriptions appropriate. Moreover, the extent to which causal language shapes listeners' social inferences remains an important question for future work (Davis et al., 2025; Kirfel et al., 2022).

In the current work, we focus on a small set of causal expressions. But natural language is flexible, allowing for diverse constructions that can express subtle distinctions in causal relations, responsibility, and intentionality. For example, our expressions only described what *did* happen, and not alternative ways of mentioning what did not (e.g., "The wizard [prevented/blocked/saved] the farmer from getting the apple."). Moving beyond these relatively simple scenarios to capture the richness of real-world discourse will involve exploring how people produce and interpret open-ended causal descriptions in naturalistic contexts. Future work will explore how our model generalizes to a broader lexicon and more complex scenarios, and how integrating dynamic social goals and their effects on speaker choices and listener inferences can advance computational models of language and social cognition.

Acknowledgments

TG was supported by grants from the Stanford Institute for Human-Centered Artificial Intelligence (HAI) and from the Cooperative AI Foundation.

References

Baglini, R., & Siegal, E. A. B.-A. (2021). Modelling linguistic causation. *manuscript, Aarhus University and Hebrew University of Jerusalem*.

Beller, A., & Gerstenberg, T. (2025). Causation, meaning, and communication. *Psychological Review*.

Bergen, L., Levy, R., & Goodman, N. (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, *9*, 20:1–91.

Cao, A., Geiger, A., Kreiss, E., Icard, T., & Gerstenberg, T. (2023). A semantics for causing, enabling, and preventing verbs using structural causal models. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45).

Davis, Z. J., Allen, K. R., Kleiman-Weiner, M., Jara-Ettinger, J., & Gerstenberg, T. (2025). Inference from social evaluation. *Journal of Personality and Social Psychology*.

Degen, J. (2023). The rational speech act framework. *Annual Review of Linguistics*, *9*, 519–540.

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford review*, *5*.

Foot, P. (2002). Killing and letting die. In P. Foot (Ed.), *Moral dilemmas*. Oxford University Press UK.

Frank, M. C., & Goodman, N. D. (2012). Predicting Pragmatic Reasoning in Language Games. *Science*, *336*(6084), 998–998.

Goldvarg, E., & Johnson-Laird, P. N. (2001). Naive causality: A mental model theory of causal meaning and reasoning. *Cognitive Science*, *25*(4), 565–610.

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

Kirfel, L., Icard, T., & Gerstenberg, T. (2022). Inference from explanation. *Journal of Experimental Psychology: General*, *151*(7), 1481–1501.

Macuch Silva, V., Lorson, A., Franke, M., Cummins, C., & Winter, B. (2024). Strategic use of english quantifiers in the reporting of quantitative information. *Discourse Processes*, *61*(10), 498–523.

McDermott, M. (1995). Redundant causation. *British Journal for the Philosophy of Science*, *46*, 523–544.

McGrath, S. (2003). Causation and the making/allowing distinction. *Philosophical Studies*, *114*(1), 81–106.

McMahan, J. (1993). Killing, letting die, and withdrawing aid. *Ethics*, 250–279.

Nadathur, P., & Lauer, S. (2020). Causal necessity, causal sufficiency, and the implications of causative verbs. *Glossa: A journal of general linguistics*, *5*(1).

Rogers, T., Zeckhauser, R., Gino, F., Norton, M. I., & Schweitzer, M. E. (2017). Artful paltering: The risks and rewards of using truthful statements to mislead others. *Journal of Personality and Social Psychology*, *112*(3), 456.

Sloman, S. A., Barbey, A. K., & Hotaling, J. M. (2009). A causal model theory of the meaning of cause, enable, and prevent. *Cognitive Science*, *33*(1), 21–50.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, *12*(1), 49–100.

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, *59*(2), 204–217.

Wolff, P. (2003). Direct causation in the linguistic coding and individuation of causal events. *Cognition*, *88*(1), 1–48.

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *136*(1), 82–111.