The language of causation

Ari Beller*, Erin Bennett* & Tobias Gerstenberg
{abeller, erindb, gerstenberg}@stanford.edu
Department of Psychology, Stanford University
450 Jane Stanford Way, Stanford, CA, USA, 94305
* equal contribution

Abstract

People use varied language to express their causal understanding of the world. But how does that language map onto people's underlying representations, and how do people choose between competing ways to best describe what happened? In this paper we develop a model that integrates computational tools for causal judgment and pragmatic inference to address these questions. The model has three components: a causal inference component which computes counterfactual simulations that capture whether and how a candidate cause made a difference to the outcome, a literal semantics that maps the outcome of these counterfactual simulations onto different causal expressions (such as "caused", "enabled", "affected", or "made no difference"), and a pragmatics component that considers how informative each causal expression would be for figuring out what happened. We test our model in an experiment that asks participants to select which expression best describes what happened in video clips depicting physical interactions.

Keywords: causality; language; counterfactuals; pragmatics; intuitive physics.

Introduction

The words we use to describe what happened matter. Hearing that "Tom killed Bill" elicits a different mental model of what happened than hearing that "Tom caused Bill to die" does (Freitas, DeScioli, Nemirow, Massenkoff, & Pinker, 2017; Niemi, Hartshorne, Gerstenberg, Stanley, & Young, 2020; Thomson, 1976). The question of how we express our knowledge of what happened in words, and how we infer what happened based on the words we hear, has a long and deep tradition in philosophy, linguistics, and cognitive science (Pinker, 2007).

In cognitive science, most work on causal cognition has focused on building models that capture the notions of "cause" and "prevent" (Waldmann, 2017). However, some attempts have also been made to uncover the differences between causal verbs such as "cause" and "enable". For example, Cheng and Novick (1991) argue that these verbs map onto different patterns of co-variation. Sloman, Barbey, and Hotaling (2009), in contrast, propose that their meanings are best understood as mapping onto qualitatively different causal models. Whereas "A causes B" means that there is a causal link from A to B, "A enables B" means that A is necessary for B to happen, and that there exists an alternative cause to B. Goldvarg and Johnson-Laird (2001) develop an account based on mental model theory in which "cause" and "allow" map onto different possible cause-effect pairs (see also Khemlani, Wasylyshyn, Briggs, & Bello, 2018). What all these accounts so far have in common is that they construe causal relationships in terms of some notion of probabilistic, counterfactual, or logical dependence.

Wolff (2007) developed a different framework for thinking about causal expressions such as "cause", "enable", "despite", and "prevent" that is based on Talmy's (1988) theory of force dynamics (see also Wolff, Barbey, & Hausknecht, 2010). According to the force dynamics model, causal expressions map onto configurations of force vectors (Figure 1). For example, the causal construction "A caused P to reach E" maps onto a configuration in which P's initial force didn't point toward the endstate E, and A's force combined with P's such that the resulting force R led P to reach the endstate. In contrast, the construction "A enabled P to reach E" implies that P's force vector *already pointed toward* the endstate, and A's force combined with P's such that it reached the endstate. The force dynamics model accurately predicts participants' modal selection of which expression best captures what happened in a variety of different video clips (Wolff, 2007). However, the force dynamics model also has some limitations. For example, it doesn't yield quantitative predictions about how well a particular expression captures what happened, but instead relies on a qualitative mapping between situations and causal expressions. Rather than a distribution over response options, it predicts a single response. Also, as we will see below, when people are provided with other alternative expressions (such as "affected", or "made no difference") they don't choose "enabled" for the predicted force configuration.

In this paper, we propose a new model of the meaning and use of causal expressions. Our model builds on the counterfactual simulation model (CSM) of causal judgment developed by Gerstenberg, Goodman, Lagnado, and Tenenbaum (2015, 2020). Gerstenberg et al. tested their model on dynamic physical interactions similar to the diagrams shown at the top of Figure 2. In their experiment, participants judged

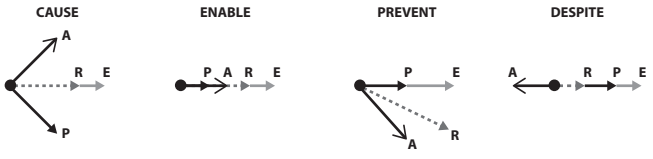

Figure 1: Analysis of different causal expressions in terms of configurations of forces. P = patient force, A = agent force, R = resulting force, E = endstate.

to what extent balls A and B were responsible for ball E's going through (or missing) the gate. The CSM predicts that people's causal judgments are sensitive to different aspects of causation that capture the extent to which the cause made a difference to *how* and *whether* the outcome happened. These aspects of causation are defined in terms of counterfactual contrasts operating over an intuitive understanding of the physical domain. People use their intuitive understanding of physics to mentally simulate how things could have turned out differently (cf. Gerstenberg & Tenenbaum, 2017; Ullman, Spelke, Battaglia, & Tenenbaum, 2017), and the result of these counterfactual simulations informs their causal judgments (Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017).

The model we develop here predicts people's use of the causal expressions "caused", "enabled", "affected", and "made no difference". The model has three components: a *causal inference component* that computes the different aspects of causation according to the CSM, a *semantics component* that defines a logical mapping from aspects of causation to causal expressions, and a *pragmatics component* that takes into account how informative each expression would be about what happened (Goodman & Frank, 2016).

The paper is organized as follows. We first describe our model and its three components by illustrating how it applies to some example cases. We then present an experiment that tests the model, as well as alternative models, on a challenging set of video clips. We conclude by discussing the model's limitations that suggest important avenues for future research.

Model

We discuss the three components of our model in turn: *causal inference*, *model semantics*, and *pragmatics*.

Causal inference

The causal inference component of our model builds on the CSM (Gerstenberg et al., 2015, 2020), which postulates that causal judgments are sensitive to different aspects of causation including how-causation, whether-causation, and sufficient-causation. We briefly describe each aspect here using clips participants viewed in our experiment (see Figure 2). Table 1a shows the aspect values for clips 1–4.

whether-causation To test for whether-causation $\mathcal{W}$, the model computes the probability that the counterfactual outcome e' in scenario S would have been different from what actually happened e, if the candidate cause A had been *removed*.

$$\mathcal{W}(A \rightarrow e) = P(e' \neq e | S, remove(A))$$

For example, in clip 1, ball A is a whether-cause of ball B's going through the gate (see Figure 2's caption for a brief description of each clip). If ball A had been removed from the scene, then ball B wouldn't have gone through the gate (because it was initially at rest). While it is clear in clip 1 that ball A was a whether-cause, in clip 8, it is less clear whether ball B would have gone through the gate if ball A had been removed. The model captures this uncertainty by running noisy counterfactual simulations. In simulating the counterfactual, B's movements are perturbed from the point at which the collision with ball A would have happened onward. By recording the proportion of cases in which the outcome would have been different from what actually happened in these noisy samples, the model computes the probability that A was a whether-cause of e. Whether-causation captures the extent to which the presence of the candidate cause was necessary for the outcome to come about.

Table 1: Model derivation for clips 1–4 shown in Figure 2.

a) Aspect Values

Clip	1	2	3	4
Whether	1.00	1.00	0.00	0.00
How	1.00	0.00	1.00	0.00
Sufficient	1.00	1.00	0.00	0.00

b) Semantic Values

Clip	1	2	3	4
No Difference	0.00	0.00	0.80	1.00
Affected	1.00	0.00	1.00	0.00
Enabled	1.00	1.00	0.00	0.00
Caused	1.00	0.00	0.00	0.00

c) Literal Listener Distributions

Clip	1	2	3	4
No Difference	0.00	0.00	0.44	0.56
Affected	0.50	0.00	0.50	0.00
Enabled	0.50	0.50	0.00	0.00
Caused	1.00	0.00	0.00	0.00

d) Speaker Distributions

Clip	1	2	3	4
No Difference	0.00	0.00	0.47	1.00
Affected	0.25	0.00	0.53	0.00
Enabled	0.25	1.00	0.00	0.00
Caused	0.50	0.00	0.00	0.00

how-causation For how-causation $\mathcal{H}$, the model computes whether the fine-grained counterfactual outcome $\Delta e'$ in scenario S would have been different from what actually happened Δe, if the candidate cause A had been *changed*.

$$\mathcal{H}(A \rightarrow \Delta e) = P(\Delta e' \neq \Delta e | S, change(A))$$

This test captures whether A made a difference to how the outcome came about (cf. Lewis, 2000; Woodward, 2011). The outcome event Δe is construed at a finer level of granularity, with information about the time and space at which the event happened. The *change()* operation is implemented as a small perturbation to ball A's initial position. For example, in clip 3 ball A knocks into ball B which was already headed toward the gate. Here, ball A is a how-cause of B's going through the gate (the outcome would have been different if ball A's initial position had been slightly perturbed) but not a whether-cause. In clip 2, ball A knocks the box out of the way so that ball B can go through the gate. Here, ball A is a whether-cause (B would not have gone through the gate if

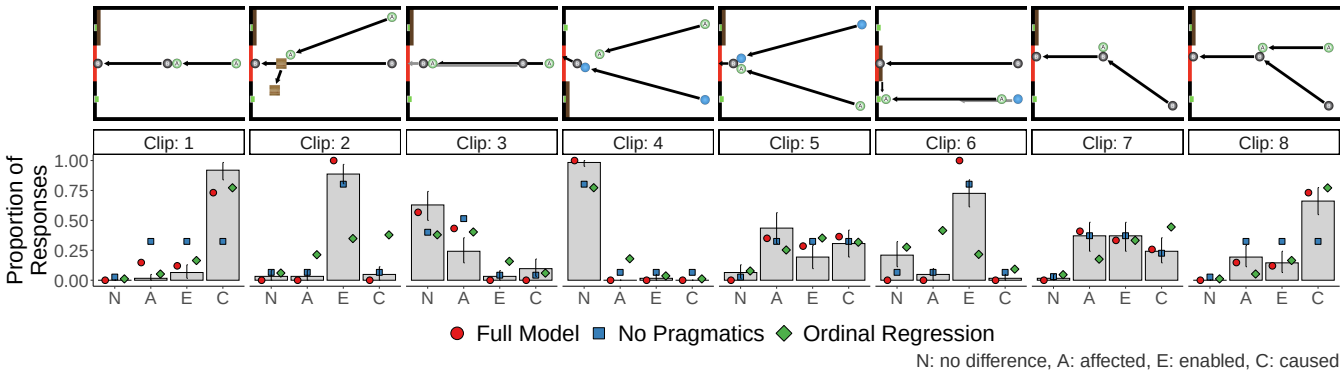

Figure 2: **Results for a selection of clips.** The clips above are an illustrative subset of the complete set of 30 clips used in the experiment. In clip 1, ball A knocks a stationary ball B into the gate. In clip 2, ball A knocks the box out of the way, enabling ball B to go through the gate. In clip 3, ball A bounces into ball B which was already headed toward the gate. In clip 4 the blue ball knocks ball B into the gate before ball A would have knocked it in. In clip 5 both balls hit ball B simultaneously. Ball B would have gone through the gate even if only one of the balls had hit it. In clip 6, ball A hits the green button that moves the door down so ball B can go through the opened gate. The blue ball would have hit the green button a little bit later if ball A hadn't been there. Clips 7 and 8 contrast a situation where ball A is stationary with one where ball A is moving. The results panels show the probabilities with which participants' selected each of the four causal expressions for each clip (bars) together with the predictions of the different models. *Note*: Error bars are bootstrapped 95% confidence intervals.

ball A had been removed) but not a how-cause (B would have gone through the gate exactly like it did, even if A's position had been somewhat perturbed).

sufficient-causation Sufficient-causation $\mathcal{S}$ is defined as

$$\mathcal{S}(A \to e) = P(\mathcal{W}(A \to e)|S, remove(\backslash A)).$$

A is sufficient for e if A would have been a whether-cause $\mathcal{W}(A \to e)$ in a situation in which the relevant alternative causes $\backslash A$ had been removed. This notion of sufficiency is inspired by Halpern and Pearl (2005) who define a test for causation that considers not only whether a candidate cause made a difference in the actual situation, but also whether it would have made a difference in other counterfactual contingencies (see also Halpern, 2016). For example, in clip 5 ball A was sufficient but not necessary for B's going through the gate. Ball A would have been a whether-cause in the situation in which the alternative cause, the blue ball, had been removed from the scene.

Also following Halpern and Pearl (2005), we constrain sufficient-causation to check whether the relevant events in the counterfactual contingency match the events that actually happened. For example, in clip 4, ball A would have knocked ball B through the gate in the counterfactual contingency in which the blue ball had been removed from the scene. However, ball A wasn't sufficient for B's going through the gate because the relevant events in this counterfactual contingency (e.g., ball A's hitting ball B) don't match the events that actually happened. For clips 1–3, there are no relevant alternative causes, so sufficient-causation reduces to whether-causation.

Model Semantics

We define a semantics of four causal expressions, "made no difference", "affected", "enabled", and "caused", as a logical mapping from causal aspects to these expressions. Table 1b shows the semantic values for clips 1–4.

"Made no difference" Our literal semantics of "made no difference" is:

$$\text{no difference}(A \to e) = \neg\mathcal{W}(A \to e) \wedge \neg\mathcal{S}(A \to e) \wedge \neg_s\mathcal{H}(A \to \Delta e)$$

Accordingly, A made no difference to e when it wasn't a whether-cause, sufficient-cause, or how-cause. The requirement to not be a how-cause is soft (which we capture by the soft-NOT $\neg_s$). This means that there is some probability that A can be said to have made no difference to the outcome e even if it was a how-cause. For example, in clip 3, ball A was only a how-cause and neither a whether-cause nor a sufficient-cause. The semantics predicts that there is still some chance of saying that ball A made no difference to B's going through the gate. In our experiment, we fit a parameter to measure the probability of responding "made no difference" even when the candidate was a how cause.

"Affected" We define "affected" like so:

$$\text{affected}(A \to e) = \mathcal{H}(A \to e)$$

A affected the outcome e if A was a how-cause. For example, in clip 7 although it is unclear if ball A was a whether-cause of B's going through the gate (B might have gone through even if ball A had been removed from the scene), it is clear that ball A affected how ball B went through the gate.

"Enabled" We define "enabled" as:

$$\text{enabled}(A \to e) = \mathcal{W}(A \to e) \vee \mathcal{S}(A \to e)$$

For A to have enabled e it must have either been a whether-cause, a sufficient-cause, or both. For example, in clip 2, ball

A enabled B's going through the gate. It was both a whether-cause and a sufficient-cause (because there were no alternative causes) of the outcome. Clip 6 demonstrates that considering sufficient-causation is important. Here, ball A hits the button that opens the door to the gate shortly before the blue ball would have hit the button. Ball A wasn't a whether-cause in this case, but it still feels right to say that ball A "enabled" B's going through the gate.

"Caused" We define "caused" as:

$$\text{caused}(A \to e) = \mathcal{H}(A \to \Delta e) \wedge (\mathcal{W}(A \to e) \vee \mathcal{S}(A \to e)) \wedge_s M(A)$$

A "caused" e when it was a how-cause of the outcome, and either a whether-cause or sufficient-cause (or both), and when A was moving $M(A)$. The requirement for ball A to have been moving is soft as indicated by $\wedge_s$. This means that even if ball A wasn't moving it can still be said to have caused the outcome sometimes.

The addition of the soft requirement for ball A's movement was motivated by the observation that prior movement affects people's causal intuitions (Mayrhofer & Waldmann, 2016; White, 2014). For example, clip 7 and clip 8 are identical in terms of their causal aspects. The only difference is that ball A is stationary in clip 7 and moving in clip 8. Participants were reluctant to say that ball A caused ball B to go through the gate when ball A was stationary.

We make the additional observation that the term "caused" is ambiguous. An event could merely be "a cause" of the outcome, or it could be "the cause". In clip 5, for example, both ball A and the blue ball collide with ball B simultaneously, knocking it into the gate. Though one could describe this situation by saying "ball A caused ball B to go through the exit", and a significant portion of participants made this selection in the experiment, doing so elides the equally important role of the blue ball in the observed causal event. It's true that ball A was "a cause" of the outcome here, but it would be misleading to say that it was "the cause" (since the blue ball played the exact same role in bringing about the outcome). To capture this ambiguity we normalized the semantic value of "caused", dividing by the sum of the semantic values of all candidate causes. This means that unique causes are more likely to be referred to by the "caused" expression. So as to not inflate the value of one candidate cause by adding alternative causes that didn't do anything, we assume that the presence of other candidate causes can only decrease the target cause's value but not increase it.

"caused" is the strongest expression in that it has the strictest requirements. A candidate can only be said to have "caused" the outcome if it made a difference to how it came about, and if it was either necessary or sufficient (or both). It is further restricted as a concept by the softer constraints that a cause must be moving and that it be unique.

Pragmatics

We use a rational speech acts (RSA) model of pragmatic reasoning to predict the probability with which a speaker would select each causal expression to describe a clip (Frank & Goodman, 2012; Goodman & Frank, 2016). The speaker chooses a causal expression by reasoning about what a listener would infer based on the utterance. This is implemented via recursive reasoning that starts with a literal listener.

The Literal Listener $L0$ updates his uniform prior beliefs about which video v the speaker refers to, conditioned on a given utterance u being literally true of that video. Formally,

$$P_{L0}(v|u) \propto P(u \text{ is true of } v) \cdot P(v),$$

An utterance is true of a video with some probability based on the semantics of the utterance and the joint distribution over aspect values.

In modeling our experiments, the literal listener is defined over the full set of video stimuli. For the purpose of demonstration, we assume that clips 1–4 in Figure 2 are the only possible videos that the speaker could be referring to. When the literal listener hears the utterance "ball A affected ball B's going through the gate", he can rule out clips 2 and 4 (since it's literally false that ball A affected ball B in these cases) but considers clips 1 and 3 to be equally likely (since "affect" is true in these clips). Assuming that each clip is equally likely a priori, the literal listener's beliefs over the different videos $P_{L0}(v|u)$ can be computed by normalizing the semantic values in Table 1b across each row to turn it into a valid conditional distribution over videos given each utterance (see Table 1c).

Notice that even though the *semantic* value for "enabled", "affected", and "caused" are all 1 for clip 1, the Literal Listener is less likely to pick out clip 1 when hearing "enabled" or "affected" compared to "caused". This is because there are fewer clips for which the more restrictive "caused" is true, and therefore that utterance is more *informative*.

We model the speaker S as soft-maximizing (with optimality parameter λ) the tradeoff between informativity to a listener and the cost c of an utterance.

$$P_S(u|v) \propto \exp(\lambda(\log P_L(v|u) - c_u))$$

The informativity is based on the surprisal that a listener would experience at finding out which video the speaker referred to. The higher probability the listener assigns to the correct video, the higher the informativity. The optimality parameter λ controls the "peakiness" of the speaker distribution. At $\lambda = 1$, the distribution is unchanged, but as λ increases, the mass of the distribution collects around the most probable response. As λ decreases, the distribution becomes increasingly flat, becoming uniform at $\lambda = 0$.

The speaker distribution in Table 1d shows the distribution over utterances for a first-level speaker reasoning about a literal listener for each of the sample clips, given $\lambda = 1.0$ and equal cost for each utterance. To compute this distribution, we multiply the Literal Listener distribution from Table 1c by a prior on utterances (in this case equal) and then we normalize again, this time over columns . We model participants with a second-level speaker, who reasons about the informativity to a *pragmatic* listener $L1$.

Experiment

We tested our model by presenting participants video clips like the ones shown in Figure 2, and asking them to select which expression best describes what happened.

Methods

Materials We created 30 video clips, including the examples from Figure 2. All stimuli included the labeled billiard balls A and B, a door, and two buttons that controlled the door's movement. Some of the stimuli also included a third (blue) ball or a box. We created this set of video clips to capture a wide range of causal interactions that are reflected in the different causal aspects.[1]

Participants We recruited 64 participants ($M_{age} = 35$, $SD_{age} = 8.24$, 19 female) online via Mechanical Turk using Psiturk (Gureckis et al., 2016). We removed two participants who failed to select "no difference" on an attention check video in which ball A unambiguously had no causal relation to the outcome. The experiment took 25 minutes and each participant was paid $3.67.

Procedure Participants received instructions and had to successfully complete a set of comprehension check questions before being able to proceed to the test phase of the experiment. In the test phase, participants viewed thirty stimulus clips plus one attention check video. The clip order was randomized between participants. Each participant watched each video at least twice before choosing one of four answers to the prompt: "Which of the following sentences best describes the clip?": (1) "Ball A **caused** ball B to go through the red exit.", (2) "Ball A **enabled** ball B to go through the red exit." (3) "Ball A **affected** ball B's going through the red exit.", or (4) "Ball A **made no difference to** ball B's going through the red exit." The order of the first three descriptions was randomized between participants, but the description with "made no difference" always came last. Participants were allowed to replay the video as many times as they liked.

Analysis

The *causal inference component* of the model has one free parameter which determines how much noise is added to ball B's motion in the counterfactual simulations for whether-causation and sufficient-causation, capturing participants' uncertainty about what would have happened. The *semantics component* has two free parameters. One for the soft-NOT in the definition of "no difference", and one for the soft-AND in the definition of "caused". The *pragmatics component* of the model has one free parameter λ to determine the speaker optimality. We fit these parameters via a grid-search minimizing squared error between aggregated participant responses and model predictions. We found an optimal value of 0.9 for uncertainty noise, 0.5 for the soft-NOT, 0.4 for the soft-AND, and 1.5 for speaker optimality.

[1]The materials for this project may be accessed here: https://github.com/cicl-stanford/causal_language_public

Alternative models We compare our model to two alternatives: a lesioned version of our full model that removes the pragmatics component, and a Bayesian ordinal regression that directly maps from aspect values to utterance selections.

No Pragmatics This model removes the RSA part of the full model, and predicts selections based on a softmax function on the semantic values. While this model retains the semantic assumptions about the mapping between causal aspects and expressions, it does not consider how informative different utterances are. Instead, it predicts participants' responses by computing a soft-max function over the semantic values (as shown in Table 1b). We found that normalizing the semantic values for "caused" hurt this model's performance, and so our reported "No Pragmatics" doesn't consider uniqueness.

Ordinal Regression We fit a Bayesian ordinal regression with coefficients for each of the causal aspects, the movement feature, and random intercepts for each participant. This model assumes the following ordering of the expressions (from weakest to strongest): "made no difference", "affected", "enabled", "caused". Unlike the other models, the regression makes no assumptions about a logical mapping from causal aspects to expressions and instead finds a linear mapping that best explains the data.

Results & Discussion

Figure 2 shows participants' selections and model fits for a subset of the clips. The full model does a good job of capturing participants' responses relative to the alternative models. In clip 1, the "No Pragmatics" model cannot distinguish between "affected", "enabled", or "caused", since all of these expressions are equally true, but the full model prefers "caused" due to its being the most informative utterance. A similar effect of informativity can be observed in clip 8. The ordinal regression works fine for clip 1 but struggles in other cases. In clip 2, the regression predicts "caused" would be preferred whereas most participants selected "enabled". In clip 6 the regression predicts "affected", which was almost never selected by participants. Interestingly, in clip 3 which maps onto the force vector configuration for "enable" (see Figure 1) according to the force dynamics theory (Wolff, 2007), participants modal response was to say that ball A "made no difference" to B's going through the gate, and some participants selected "affected". Almost no one selected "enabled" in this case.

Figure 3 shows scatter plots of model predictions and aggregated participants' responses for the full set of clips. The full model's predictions correlate best with participants' responses and show the lowest error, followed by the "No Pragmatics" model, and lastly the ordinal regression. This pattern of decreasing performance with each successive lesion suggests that the addition of both the explicit semantics, and the pragmatic comparison provide important contributions to the overall model performance. To further assess model fit, we performed 100 split-half cross validation runs for each model, splitting the data by trials. The full model achieves a correlation of $r = 0.91$ $[0.83, 0.94]$ (median [5%, 95% percentile]),

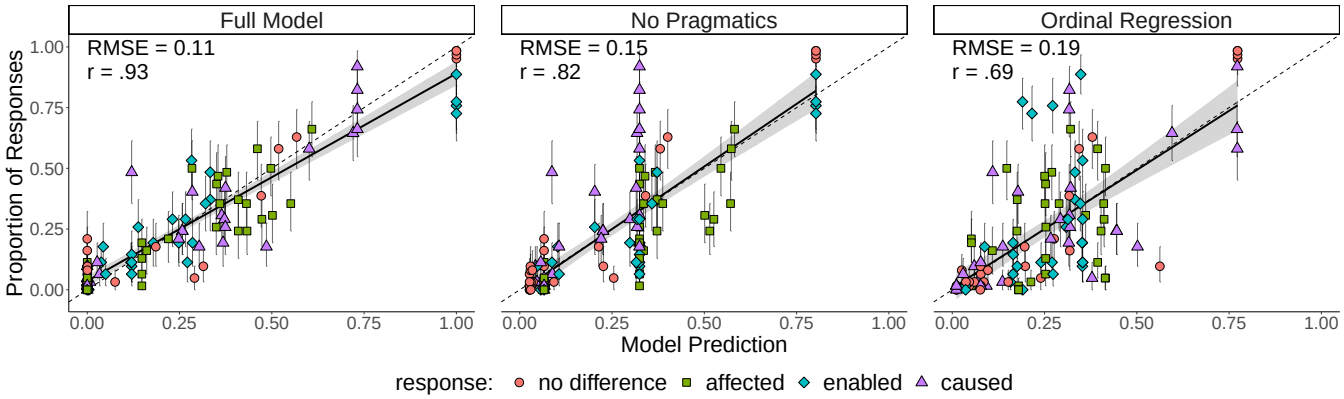

Figure 3: **Experiment model fits.** Scatter plots of model predictions (x-axis) and proportion of participants selections of each causal verb (y-axis) for the full model (left), the model without pragmatics (middle), and the ordinal regression. Error bars are bootstrapped 95% confidence intervals.

outperforming "No Pragmatics" ($r = 0.82$ $[0.73, 0.90]$), and the ordinal regression ($r = 0.59$ $[0.30, 0.74]$).

To what extent can the effects of pragmatic inference be captured by more fine-grained semantic distinctions? Participants almost unanimously endorsed "caused" in clip 1 and "enabled" in clip 2, which might give the impression that participants don't think that there are multiple true utterances in these clips. Can we improve or at least match the performance of the full model by encoding more in the semantics? We modified the semantics of "enabled" to explicitly include a negation of how-cause thus rendering "cause" and "enable" mutually exclusive. The resulting model has a lower performance overall ($r = 0.87, \text{RMSE} = 0.16$). It struggles with cases like 5 and 8 where participants endorse a wider range of descriptions. But this observation does raise an interesting question. To what extent are participant's selections in this task a reflection of their semantic understanding of the causal expressions versus a reflection of pragmatic communicative pressures in context?

General Discussion

In this paper, we developed a novel model of the meaning and use of different causal expressions. The model builds on a counterfactual simulation model to compute different aspects that capture the way a candidate cause made a difference to the outcome (Gerstenberg et al., 2015, 2020). It uses these aspects of causation to define a literal semantics of a set of causal expressions including "caused", "enabled", "affected", and "made no difference". Finally, it uses pragmatic inference to predict language use by taking into account which expression would be most informative about what happened. The model accurately captures what expressions participants select for a range of video clips depicting dynamic physical interactions. An important contribution of this work is the integration of a model of people's causal knowledge with tools for modeling pragmatic linguistic communication (Goodman & Frank, 2016). The comparison of our full model with lesioned models shows that both the semantic and the pragmatic components of the model are critical.

The model developed here makes a first step toward a more complete theory of causal language use, but is still limited in important ways. For example, the test for sufficient-causation includes a quantifier over relevant alternative causes, as well as relevant causal events. Defining in a more principled way what should be included in these sets (e.g. should the box be treated as an alternative cause?) is an open question for future research (Hesslow, 1988).

Our full model includes a softening parameter in the semantics for "made no difference" that captures the ambiguity between made no differnce to *how* versus *whether* the outcome occurred. One way to derive this softening in a more principled way might be by assuming uncertainty in the semantics, which can then be resolved by pragmatic inference (cf. Bergen, Levy, & Goodman, 2016).

Our literal semantics of "caused" includes the movement feature which captures people's reluctance to use the expression "caused" for stationary causes. However, the inclusion of this feature seems somewhat ad-hoc and it would be nice to derive it in a more principled way from the CSM. One possibility is to strengthen the definition of sufficiency. Only moving balls can be strongly sufficient-causes in our setting – only moving balls have the capacity to bring about an outcome without help of other additional causes (like in clip 1). A stationary cause can never be strongly sufficient in that sense – it always requires the existence of some other cause to make the outcome happen. For example, in clip 7, some other cause must be responsible for ball B's initial movement.

We focused here on a relatively small set of causal expressions. In future work, we aim to expand our model to capture additional causal expressions such as "helped", "allowed", "let", and "made" (cf. Lauer & Nadathur, accepted). Eventually, we would like to develop the model to produce written explanations to questions such as "Why did ball B go through the gate?" In tandem with such an account of explanation generation, we aim to develop a model of explanation understanding that can infer from causal explanations what happened in a video.

Acknowledgments

We thank Beth Levin, Dan Lassiter, Elisa Kreiss, Robert Hawkins, Thomas Icard, the Causality in Cognition Lab, and anonymous reviewers for helpful discussion and advice.

References

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

Cheng, P. W., & Novick, L. R. (1991). Causes versus enabling conditions. *Cognition*, *40*, 83–120.

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

Freitas, J. D., DeScioli, P., Nemirow, J., Massenkoff, M., & Pinker, S. (2017). Kill or die: Moral judgment alters linguistic coding of causality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In *Cogsci.*

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2020). A counterfactual simulation model of causal judgment. *PsyArXiv*. (`https://psyarxiv.com/7zj94/`)

Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017, oct). Eye-tracking causality. *Psychological Science*, *28*(12), 1731–1744.

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. In M. Waldmannn (Ed.), *Oxford handbook of causal reasoning* (pp. 515–548). Oxford University Press.

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

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

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., … Chan, P. (2016). psiturk: An open-source framework for conducting replicable behavioral experiments online. *Behavior research methods*, *48*(3), 829–842.

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

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

Hesslow, G. (1988). The problem of causal selection. In D. J. Hilton (Ed.), *Contemporary science and natural explanation: Commonsense conceptions of causality* (pp. 11–32). Brighton, UK: Harvester Press.

Khemlani, S., Wasylyshyn, C., Briggs, G., & Bello, P. (2018). Mental models and omissive causation. *Memory & cognition*, *46*(8), 1344–1359.

Lauer, S., & Nadathur, P. (accepted). Causal necessity, causal sufficiency, and the implications of causative verbs. *Glosssa*.

Lewis, D. (2000). Causation as influence. *The Journal of Philosophy*, *97*(4), 182–197.

Mayrhofer, R., & Waldmann, M. R. (2016). Causal agency and the perception of force. *Psychonomic Bulletin & Review*, *23*(3), 789–796.

Niemi, L., Hartshorne, J., Gerstenberg, T., Stanley, M., & Young, L. (2020, June). Moral Values Reveal the Causality Implicit in Verb Meaning. *Cognitive Science*, *44*(6).

Pinker, S. (2007). *The stuff of thought: Language as a window into human nature*. Penguin.

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.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017, sep). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, *21*(9), 649–665.

Waldmann, M. R. (Ed.). (2017). *The oxford handbook of causal reasoning*. Oxford University Press.

White, P. A. (2014). Singular clues to causality and their use in human causal judgment. *Cognitive Science*, *38*(1), 38–75.

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

Wolff, P., Barbey, A. K., & Hausknecht, M. (2010). For want of a nail: How absences cause events. *Journal of Experimental Psychology: General*, *139*(2), 191–221.

Woodward, J. (2011). Mechanisms revisited. *Synthese*, *183*(3), 409–427.