

A counterfactual simulation model of causal language

Ari Beller

Department of Psychology, Stanford University

Tobias Gerstenberg

Department of Psychology, Stanford University

June 30, 2023

Abstract

The words we use to describe what happened shape the story a listener imagines. How do speakers choose what causal expression to use? How does that impact what listeners infer about what happened? In this paper, we develop a computational model of how people use the causal expressions “caused”, “enabled”, “affected”, and “made no difference”. The model first builds a causal representation of what happened. By running counterfactual simulations, the model computes causal aspects that capture the different ways in which a candidate cause made a difference to the outcome. Logical combinations of these aspects define a semantics for the different causal expressions. The model then uses pragmatic inference favoring informative utterances to decide what word to use in context. We test our model in a series of experiments. In a set of psycholinguistic studies, we verify semantic and pragmatic assumptions of our model. We show that the causal expressions exist on a hierarchy of informativeness, and that participants draw informative pragmatic inferences in line with this scale. In the next two studies, we demonstrate that our model quantitatively fits participant behavior in a speaker task and a listener task involving dynamic physical scenarios. We compare our model to two lesioned alternatives, one which removes the pragmatic inference component, and another which additionally removes the semantics of the causal expressions. Our full model better accounts for participants’ behavior than both alternatives, suggesting that causal knowledge, semantics, and pragmatics are all important for understanding how people produce and comprehend causal language.

Keywords: causality; counterfactuals; mental simulation; intuitive physics; language; semantics; pragmatics.

Corresponding author: Tobias Gerstenberg, Stanford University, Department of Psychology, 450 Jane Stanford Way, Bldg 420, Stanford, CA 94305, Email: gerstenberg@stanford.edu. All the data, study materials, and analysis code are available here: https://github.com/cic1-stanford/causal_language

Introduction

The words we use matter. When you hear someone say “Tom killed Bill”, an image of a murderous scene pops into your mind. If you hear instead that “Tom caused Bill to die”, you might imagine a different scenario, one that’s less direct and murderous than the first. Both of these sentences attribute a causal role to Tom in Bill’s death, but the subtle differences in phrasing amount to worlds of difference in meaning. In everyday communication, causation enters into people’s language in all kinds of innocuous yet impactful ways. While it is easy to miss the significance of these linguistic choices in the moment, their variety and flexibility supports our capacity to effortlessly convey and comprehend causal stories that are both complex and specific.

Studying the language of causation is a multi-disciplinary endeavor. Linguists aim to understand how people talk about cause and effect (Aronson, 1971; Garvey & Caramazza, 1974; Hobbs, 2005; Levin & Hovav, 1994; Shibatani, 1976; Talmi, 1988) and in doing so often draw inspiration from the philosophical literature (Dowe, 2000; Hall, 2004; Lewis, 1973). Psychologists study how people represent causal relationships and how they choose specific causal expressions on the basis of that knowledge (Cheng & Novick, 1991; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021; Rose, Sievers, & Nichols, 2021; Wolff, 2007). In this paper we develop the *counterfactual simulation model of causal language* which incorporates insights from these three disciplines. Our model combines a psychological approach for representing causation that is grounded in the philosophical literature on causation, with linguistic tools for modeling meaning and pragmatic communication. Our model builds on the Counterfactual Simulation Model (CSM, Gerstenberg et al., 2021), a quantitative model of how people make causal judgments. We use the CSM to define a semantics for various causal expressions, and use the rational speech acts model (Degen, 2023; Frank & Goodman, 2012) to capture pragmatic inferences in communication. The combination of these techniques helps us shed light on the interaction of language and thought in how people communicate about causality. Before describing how the model works, we discuss the relevant background literature from each of these three disciplines: philosophy, linguistics, and psychology.

The philosophy of causal language

There are two major philosophical approaches for analyzing causation: *dependence theories* and *process theories*. According to dependence theories, causality is a dependence relation between cause and effect. Dependence has been characterized in various ways. For example, in counterfactual theories, A is a cause of B if A and B happened, and it’s true that B would not have happened if A had not happened (Lewis, 1973; Mackie, 1974; Woodward, 2003). Imagine that Marco threw a stone at a window and the window broke. Here, Marco’s throwing the stone caused the window to break because both of these events happened, and because the window wouldn’t have broken had Marco not thrown the stone. On the other hand, process theories claim that causal relationships are defined by spatiotemporally continuous processes that link causes with effects (Dowe, 2000; Machamer, Darden, & Craver, 2000; Salmon, 1984). So, A is a cause of B if A transferred some property, such as energy or force, to B. Here, Marco caused the window to break because the stone transferred force to the breaking window via a spatiotemporally contiguous process.

Philosophers are not only interested in metaphysical questions about what causation is. They also care about developing theories that accord with human intuitions about causation. For these philosophers, the semantics of the word “caused” is a key topic of inquiry, and they investigate its meaning by identifying the sets of circumstances in which people believe it is appropriate to use the word (e.g. Fair, 1979; McDermott, 1995; Schaffer, 2013). For example in the dependence tradition, Halpern and Pearl (2005) develop an account of when it’s appropriate to say that “A caused B” using the framework of structural equation models (SEM), a formal language for describing causal relationships (Pearl, 2000; Spirtes, Glymour, & Scheines, 2000). Halpern and Pearl validate their approach by considering a series of counterexamples to counterfactual theories of causation from the philosophical literature (see also Hall, 2007; Halpern, 2016; Hitchcock, 2009; Hitchcock & Knobe, 2009; Livengood, 2011; Woodward, 2021). They show that, for an appropriate SEM, their definition identifies those variables as causes that match common intuitions. For example, one of the cases they consider is an instance of causal preemption (see Hall, 2004; Paul & Hall, 2013). In this situation, Billy and Suzy both throw stones at a bottle. Suzy’s stone gets there first and the bottle shatters, but even if Suzy hadn’t thrown hers, Billy’s throw was right on target such that the bottle would have broken anyway. People generally believe that Suzy’s stone was the cause of the bottle shattering, but a naive counterfactual analysis says that Suzy’s stone didn’t cause the bottle to shatter, because even if she hadn’t thrown it, the bottle still would have shattered anyhow (due to Billy’s stone). Halpern and Pearl accommodate these intuitions by modifying the simple counterfactual definition to consider not only what happened in the actual situation, but also what could have happened in relevant counterfactual contingencies. A variable can qualify as a cause of the outcome so long as there is an admissible contingency in which the effect counterfactually depends on the cause.

Other philosophers are even more explicit about the connection between analyzing causation and accounting for people’s use of causal language. Working in the process tradition, Fair (1979) states that a central mystery in understanding causation is accounting for the broad consistency in people’s agreement to causal statements. According to Fair, solving this mystery requires “stating explicit truth-conditions for simple declarative sentences containing the word, ‘cause,’” (Fair, 1979, pg. 220). Fair doesn’t offer these truth-conditions himself, but he suggests a common feature that people are sensitive to when they interpret situations as causal. According to Fair, causation is reducible to energy-momentum flow from cause to effect. In a game of pool, a cue-ball causes an eight-ball to sink because the cue-ball transferred energy to the eight-ball upon collision, and that energy carried the eight-ball to into the pocket (but see Hitchcock, 1995).

Philosophers not only analyze the meaning of “cause”, they also care about the differences between various causal expressions such as “cause” versus “affect” (McDermott, 1995). The subtle differences between causal expressions have animated debates in moral philosophy. For example, philosophers differentiate “killing” from “letting die” on the basis of causal structure (Foot, 1967; Malm, 1989; McGrath, 2003; McMahan, 1993; Thomson, 1976b). The distinctions that drive this discussion rest on the causal status of omisive causes (causing something to happen by “not preventing” it from happening) and sufficient causes (Gerstenberg & Stephan, 2021; Mackie, 1974; McGrath, 2003; Schaffer, 2000). Whether we say “killing” or “letting die” has practical significance in bioethics, and

the meanings of these words are wrapped up in broader debates about end-of-life care in medicine (Rodríguez-Arias, Rodriguez Lopez, Monasterio-Astobiza, & Hannikainen, 2020), a woman's right to an abortion (Thomson, 1976a), and the moral obligation to vaccinate (Flanigan, 2014).

The linguistics of causal language

Linguists have studied the many ways in which causality reveals itself in language both implicitly (e.g. Garvey & Caramazza, 1974; Hartshorne, 2013; Niemi, Hartshorne, Gerstenberg, Stanley, & Young, 2020) and explicitly (e.g. Kaufmann, 2013; Talmey, 1988). Causative constructions come in different varieties. Two of the most common in English are periphrastic causatives, where the causal meaning is expressed in a domain-general verb (e.g., “Jane **caused** the ice to break”), and lexical causatives, where the causal meaning is embedded in a domain-specific verb (e.g., “Jane **broke** the ice”). The semantic relationships between lexical causatives and corresponding periphrastic causatives with analogous meanings (“killed” vs. “caused to die”) is a topic of extensive discussion (Cruse, 1972; Fodor, 1970; Shibaiani, 1976; Smith, 1970; Wierzbicka, 1975). Though at first glance, lexical causatives and corresponding periphrastic paraphrases might seem to mean the same thing, linguists have noted that their meanings can come apart. Katz (1970) provides an example where a sheriff is set to duel with an outlaw. Prior to the duel, the sheriff has his gun poorly repaired by a gunsmith, such that when the sheriff ultimately faces the outlaw, the gun doesn't fire and the sheriff is shot dead. According to Katz, the gunsmith caused the sheriff to die, but he didn't kill him. The “killing” description is reserved for the outlaw who actually shot the sheriff.

A common theme emerging from this literature is that lexical causatives imply some form of direct causation over and above the general causal relationships implied by periphrastic causatives. It is appropriate to say that the outlaw killed the sheriff (because the actions of the outlaw are the direct cause of the sheriff's death), while the gunsmith merely caused the sheriff to die. Baglini and Siegal (2021) model the semantic differences between lexical and periphrastic causatives using SEMs. Baglini and Siegal argue that the different causative constructions imply different structural roles of causes cited in an underlying SEM. Periphrastic causatives (specifically the periphrastic causative “cause”) can be applied to any variable in the SEM that is a necessary condition for some observed outcome. However, to qualify as the subject of a lexical causative, a variable in the SEM must be part of a set of causal conditions that are jointly sufficient to bring about the outcome. Baglini and Siegal show that these definitions account for the intuitions of direct causation that have shaped many earlier accounts, while also addressing counterexamples to the directness criteria that have been raised more recently (see Wolff, 2003, for a review of different approaches to direct causation). In a similar vein, Nadathur and Lauer (2020) use the SEM framework to explain semantic differences between “caused” and “made”, whereby “caused” is analyzed in terms of necessity, and “made” in terms of sufficiency.

The SEM approaches to causal semantics are rooted in dependence theories of causation. Within the process theory tradition, Talmey (1988) developed the force-dynamics model for analyzing the meaning of various periphrastic causatives. A force-dynamic description identifies a focal entity called an agonist and an opposing entity called an antagonist along with their intrinsic tendencies toward action or rest, their relative strengths, and a resultant

action. For example, in a situation where a ball knocks into a table lamp and the lamp falls over, the agonist would be the table lamp which has a tendency toward rest, while the antagonist would be the ball which has a tendency toward action. In this case, the ball's tendency toward action overpowers the lamp's tendency toward rest leading to a resultant action where the lamp falls over. According to Talmy, this particular force-dynamic configuration corresponds to the prototypical causation scenario – a scenario that people would describe using the periphrastic causative “cause”.

The work discussed so far focuses on the semantics of causative constructions. Pragmatic inferences, however, also impact their interpretation (Degen, 2023; Goodman & Frank, 2016; Grice, 1989; Schaffer, 2013). McCawley (1978) suggests that principles of cooperative conversation influence people's selection of causative constructions in context. For example, the presence of alternatives affects people's inferences about the directness of causation. Upon hearing that “the gunsmith caused the sheriff to die”, a listener is likely to infer that the gunsmith's causal role was indirect because the speaker could have used the alternative “killed” to communicate a more direct causal role. However, when there is no lexical alternative, such as when “Bill caused Mary to laugh”, the periphrastic causative is acceptable regardless of whether causation was direct or indirect. McCawley argues that inferring indirect causation from periphrastic causatives isn't due to their semantics, but rather that it's a type of pragmatic inference called a conversational implicature Grice (1989). Listeners in conversation generally assume that speakers are as informative as they can be, and if they choose a less specific description (“caused to die”) when a more specific one is available (“killed”), this suggests that the more specific situation isn't true.

The psychology of causal language

Psychologists study the mental representations underlying people's choice of particular causal expressions. In the tradition of statistical dependence accounts of causality, Cheng and Novick (1991) argue that people represent causes and enabling conditions as patterns of covariation across different contexts. Goldvarg and Johnson-Laird (2001) use mental model theory (Johnson-Laird, 1989) to develop an account in which people's use of “cause” and “allow” maps onto different logical relationships about the presence and absence of cause and effect. Sloman, Barbey, and Hotaling (2009) suggest causal models (Pearl, 2000; Sloman, 2005) as an alternative underlying representation. According to their account, “A causes B” implies that the speaker believes there is a direct causal link between A and B, while “A enables B” implies that A is necessary for B *and* there exists an alternative direct cause that makes B happen. All of these accounts use modeling tools from dependence accounts of causation. They argue that causal language can be analyzed in terms of psychological models that represent statistical, logical, or structural dependence.

In the tradition of process theories of causation, Wolff (2007) developed an account that captures the meanings of different periphrastic causatives (“caused”, “enabled”, “despite”, and “prevent”) using configurations of forces (see also Wolff, Barbey, & Hausknecht, 2010). Wolff's account builds directly on Talmy's (1988) force dynamics theory. Figure 1 illustrates how different force configurations map onto different causal expressions. The force dynamics model assumes the presence of a causal agent, a causal patient, and an endstate. The model defines “cause” as a situation where the patient's initial force vector is not directed toward the endstate, but the combination of the agent and patient force

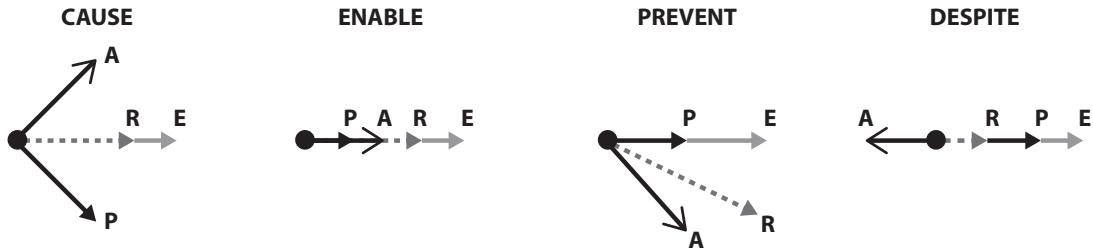
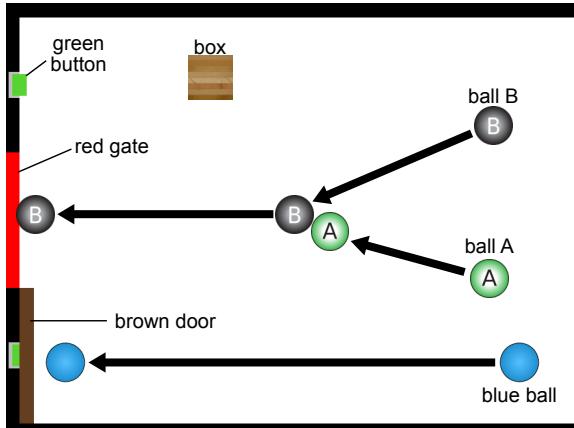


Figure 1. According to the force dynamics model (Wolff, 2007), different configurations of forces map onto different causal expressions. Each force configuration involves an agent force (A), a patient force (P), a resultant force (R), and an endstate (E).

vectors yield a resultant force vector that directs the patient toward the endstate. The model defines “enable” as a situation where the patient’s force vector is already directed toward the endstate, and the combination with the agent’s force vector yields a resultant vector that is amplified toward the direction of the endstate. The force dynamics model predicts participants’ modal selections of causal expressions across a range of animated video clips. However, it has certain limitations. Most notably, the model only predicts a qualitative mapping from force vector configurations to words. It doesn’t make quantitative predictions about the probability of choosing one expression or another.

In this paper, we build off a recent approach for modeling causal judgments: the counterfactual simulation model (CSM, Gerstenberg et al., 2021). The CSM draws on insights from both dependence and process theories of causation. Like counterfactual theories, it postulates that people make causal judgments by imagining what would have happened in counterfactual situations and comparing those counterfactual outcomes to what actually happened. Like process theories, the CSM suggests that the process by which the events took place is a crucial part of the psychological representation. Unlike earlier dependence accounts which rely on abstract dependence relations that are removed from the underlying mechanisms, in the CSM, counterfactuals are simulated in a noisy physics engine in a way that is sensitive to the details of the physical process (cf. Chater & Oaksford, 2013; Gerstenberg & Tenenbaum, 2017; Jensen, 2019). For example, consider the scenario in Figure 2. In this situation a target ball, B, enters from the right side along with a candidate cause ball, A. They collide, and ball B ends up going through the gate. To determine whether ball A caused ball B to go through the gate, the CSM runs a counterfactual simulation where ball A is removed from the scenario. If ball B does not go through the gate in this counterfactual, the CSM determines that ball A caused the outcome. Gerstenberg et al. (2021) showed that the CSM accurately captures people’s graded causal judgments across a variety of scenarios. Here, we extend the CSM to model people’s use and interpretation of several causal expressions: “caused”, “enabled”, “affected”, and “made no difference”.

The rest of the paper proceeds as follows. We begin by introducing our model, reviewing the counterfactual simulation model, and elaborating on how we link up this psychological model with tools from semantics and pragmatics to predict causal language use. We then present a series of experiments. The first is a psycholinguistics task where we validate our model’s semantic and pragmatic assumptions. The second and third experiments test the model’s quantitative predictions. We present participants with a speaker



Did ball A cause ball B to go through the gate?

Figure 2. An example of the billiard ball setting and the objects it contains. The green ball labeled A is the candidate cause, and the grey ball labeled B is the target of the causal interaction. The blue ball and the box are auxiliary objects which can influence the outcome. The brown door in front of the red gate blocks the exit. If any object contacts the green button below or above, the door will move in the direction of the button that was pressed and come to a stop once it touches the side wall.

task (Experiment 2) and a listener task (Experiment 3). We compare our model’s behavior to two alternative models that lesion different portions of the model capabilities. Our analysis reveals that the full model including causal knowledge, semantic representations, and pragmatic linguistic inference does the best job of explaining participant behavior. We close in the General Discussion by considering implications of the work and suggesting directions for future research.

A counterfactual simulation model of causal language

Our model combines causal reasoning with pragmatic communication to produce and understand causal language. The model has three components: a causality module, a semantics module, and a pragmatics module. Figure 3 illustrates how the model works. The causal knowledge module computes a causal representation of a scenario. This causal representation then feeds into the semantics module which determines which expressions are true in the scenario. Finally, on the basis of this semantics, a pragmatics module chooses an expression which is both true and informative for the given scenario. We discuss each of these components in turn and illustrate how they work via the sample cases in Figure 4.

Causality Module

The causal knowledge component of our model is based on the CSM (Gerstenberg et al., 2021). The CSM is a quantitative model predicting causal judgments in physical settings (see also Gerstenberg, 2022; Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017; Gerstenberg & Stephan, 2021; Zhou, Smith, Tenenbaum, & Gerstenberg, 2023).

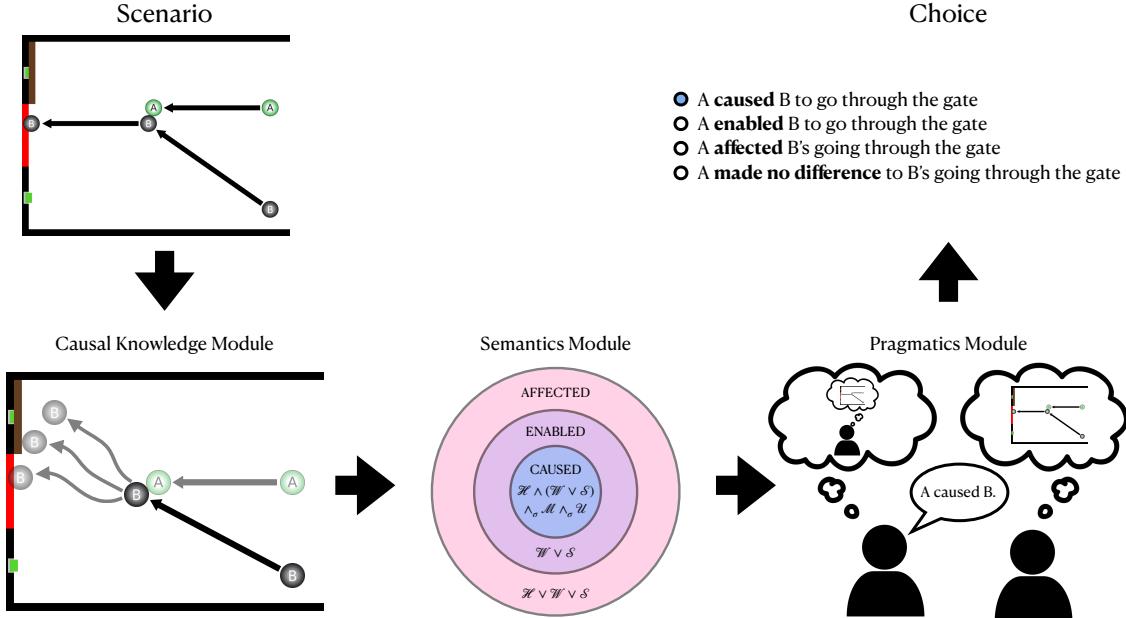


Figure 3. A schematic of the counterfactual simulation model in the speaker task. The model takes a scenario and runs different counterfactual simulations to compute several aspects of causation that capture whether and how a candidate cause made a difference to the outcome (causal knowledge module). The meaning of various causal expressions is defined through logical combinations of these aspects of causation (semantics module). The model considers both what’s true and what’s informative when deciding what expression to choose (pragmatics module).

The CSM postulates that people make causal judgments by imagining what would have happened in counterfactual situations and comparing those counterfactual outcomes to what happened in the actual situation (Lewis, 1973; Pearl, 2000). The CSM generates counterfactuals by running simulations in an intuitive physics model. Building on prior work, we model intuitive physical thought using a noisy physics simulator (Battaglia, Hamrick, & Tenenbaum, 2013; Gerstenberg & Tenenbaum, 2017; Kubricht, Holyoak, & Lu, 2017; Smith et al., in press; Ullman, Spelke, Battaglia, & Tenenbaum, 2017).

Noisy simulators allow us to capture uncertainty about what would have happened in relevant counterfactual situations. When humans simulate how a counterfactual would have played out, various sources of noise affect the accuracy of their simulations (Smith & Vul, 2013). To model this uncertainty, we inject a small amount of Gaussian noise to the direction of the velocity vectors of the objects at the point at which the counterfactual simulation diverges from what actually happened. For example, in Figure 2, if we want to simulate the counterfactual where ball A was not there, we take ball A out of the scenario and run the simulation forward. In the counterfactual, Ball B’s path diverges from what happened in the actual situation at the time point where the collision took place. At this point in the counterfactual simulation we begin applying noise to ball B’s velocity, reflecting people’s uncertainty about what would have happened. The amount of noise in the physical

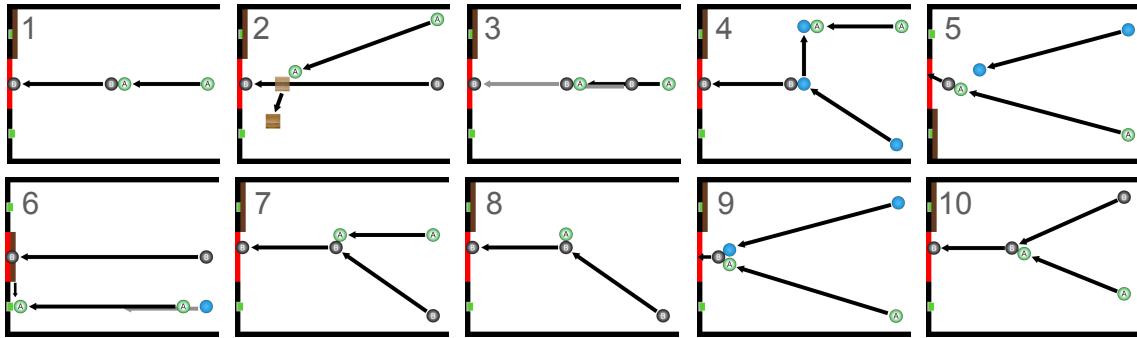


Figure 4. Sample scenarios from Experiments 2 and 3. 1) Classic Michottean case. Ball B sits in the middle of the scene and ball A comes in from the side and knocks it through the gate. 2) Ball B is headed toward the gate, and ball A knocks a box blocking its path out of the way. 3) Ball B is moving toward the gate unobstructed. Ball A comes up from behind and pushes it along, speeding it up. 4) The blue ball knocks ball B through the gate. Afterward the blue ball collides with ball A. 5) A case of causal preemption. Ball A knocks ball B through the gate shortly before the blue ball would have done the same. 6) Similar to scenario 5 but without direct contact. Here ball A pushes the button opening the gate shortly before the blue ball would have done the same. Opening the gate allows ball B to pass through the gate. 7) Ball A and ball B enter from the right side. The collision redirects ball B through the gate. 8) Similar to scenario 7 except here ball A is stationary. 9) Ball A and the blue ball collide simultaneously with ball B, pushing it through the gate. 10) Ball A collides with ball B and ball B goes through the gate. Here, it is unclear whether ball B would have gone through on its own without ball A.

simulations is determined by a free parameter of the model, θ , the standard deviation of the noise distribution.

The CSM posits that people are sensitive to multiple *aspects* of causation. These aspects represent different causal features that have been shown to affect causal judgments. The CSM computes the different aspects of causation by simulating different counterfactual possibilities. Here, we consider three aspects of causation: whether-causation, how-causation, and sufficient-causation.

whether-causation. Whether-causation corresponds to the notion of counterfactual necessity. This is the traditional counterfactual concept of causation according to which A caused B when both A and B took place and when B would not have taken place without A. The CSM evaluates whether-causation \mathcal{W} by computing 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*.

$$\mathcal{W}(A \rightarrow e) = P(e' \neq e | s, \text{remove}(A)). \quad (1)$$

In words, to test if ball A was a whether-cause of ball B going through the gate (e), we stage a counterfactual where we remove ball A from the scene and then simulate what would have happened in that counterfactual simulation. If ball B would have gone through the gate anyway, then we determine that ball A did not make a difference to whether or not the event occurred, so it is not a whether-cause. However, if ball B would not have gone through

the gate in the counterfactual, we determine that ball A was indeed a whether-cause of the outcome. This single evaluation yields a binary determination, but as we noted above, people exhibit uncertainty in their counterfactual judgments. To capture this gradation, we compute the probability that ball B would have gone through the gate in ball A’s absence by running multiple noisy simulations. The proportion of simulations in which the outcome in the counterfactual situation would have been different from what actually happened yields a graded measure of whether-causation.

For example, consider scenario 1 in Figure 4. In this scenario, ball B is stationary in the middle of the screen, and ball A comes in from the side and collides with ball B, launching it through the gate. To test for whether-causation in this situation, we remove the candidate cause (ball A) from the scene and then run multiple counterfactual simulations in its absence. In this case, it is very clear that ball B would not have gone through the gate in the counterfactual because it was stationary at the start of the scenario and only picked up momentum after the collision with ball A. After simulating multiple times, the model determines that in scenario 1, ball A has a whether-cause value of 1.0 (the counterfactual is always different from the actual outcome). While the evaluation is very clear in this scenario, it is less clear in scenario 10, for example, where ball B has its own initial momentum. If we ran noisy simulations after removing ball A from this scenario, ball B would go through the gate on some simulations and miss the gate on others. With a noise value of 1.0, the model computes a whether-cause value of $\mathcal{W}(A \rightarrow e) = 0.76$ in this scenario.

how-causation. Counterfactual necessity is an important part of the story of how people make causal judgments. However, it’s not the full story. Take for example scenario 2 where ball A knocks a box that is blocking ball B’s path out of the way and then ball B goes through the gate. This is a case of double prevention: ball A prevented the box from preventing ball B from going through the gate. Just as in scenario 1, the presence of Ball A is counterfactually necessary for ball B to go through the gate. However, prior work shows that people often rate candidate causes in double prevention scenarios as less causal than in more standard cases like scenario 1 (Gerstenberg et al., 2021; Henne & O’Neill, 2022; Lombrozo, 2010). Process theorists explain the difference between these two scenarios by appealing to the direct transfer of force from the cause to the target in scenario 1 (e.g. Wolff et al., 2010). In scenario 2 there is no direct transfer.¹

The CSM accounts incorporates information about the causal process by testing for how-causation. Testing for how-causation determines whether a candidate cause made a difference to *how* the outcome came about at a fine level of granularity. Whereas whether-causation assesses difference-making at the level of the outcome event (whether ball B went through the gate or didn’t), how-causation is sensitive to the precise details of how that event came about. A candidate cause A , is a how-cause of the fine-grained outcome Δe , if in scenario s , the fine-grained counterfactual outcome $\Delta e'$ would have been different if the candidate cause A had been *changed*:

$$\mathcal{H}(A \rightarrow \Delta e) = P(\Delta e' \neq \Delta e | s, \text{change}(A)). \quad (2)$$

We define the “fine-grained outcome” as the precise position and time at which ball B passes

¹While process theorists can account for the difference between these two cases, they struggle to account for why people feel any inclination to give a causal rating in scenario 2 where there is no direct transfer of force to the target (but see Wolff et al., 2010).

through the gate, while the *change* operator is implemented as a small perturbation to the initial position of the candidate cause. If this small perturbation leads to a difference in the final position or time at which ball B passes through the gate, the candidate cause is a how-cause.² In scenario 1, ball A is indeed a how-cause, the slight change in A results in a slight change in the fine-grained outcome. By contrast, in scenario 2, the perturbation makes no difference. Thus, the notion of how-causation helps us understand the qualitative difference between these two cases and accounts for that difference with a formal test.

sufficient-causation. Whether-causation and how-causation express much of the causally relevant information about what happened. However, these two components still fail to capture certain intuitions. A notable set of objections to counterfactual theories of causation center around cases of causal preemption (Bunzl, 1980; Hall, 2004; McDermott, 1995; Wolff, 2007). These are situations like the one presented in the introduction where Billy and Suzy throw stones at the same bottle, and Suzy’s stone hits the bottle first. Figure 4 scenario 5 depicts a billiard ball scenario with analogous causal structure. In this scenario, ball A knocks into ball B and sends it through the gate shortly before the blue ball would have done the same. Prior work shows that people judge ball A to have caused ball B to go through the gate in this situation even though ball A is not counterfactually necessary for that outcome (Gerstenberg et al., 2021). Whether-causation and how-causation alone cannot explain this pattern.

To capture people’s intuition in preemption cases, it seems necessary to include sufficiency (Beckers, 2021; Gerstenberg et al., 2021; Halpern & Pearl, 2005; Icard, Kominsky, & Knobe, 2017; Woodward, 2006). Intuitively, sufficiency tells us whether a candidate cause would have been enough to bring about the outcome “on its own”. A counterfactual test for sufficiency simulates whether the candidate cause would have been a whether-cause in the counterfactual contingency in which alternative causes had been removed. For example, if we wanted to know whether ball A was a sufficient-cause for ball B to go through the gate in scenario 5, we would first remove the alternative causes from the scene (in this case the blue ball), and then test in this contingency if ball A would have been a whether-cause. Accordingly, a candidate cause A is a sufficient-cause of outcome e in scenario s if:

$$\mathcal{S}(A \rightarrow e) = P(\mathcal{W}(A \rightarrow e)|s, \text{remove}(\setminus A)). \quad (3)$$

Here, $\text{remove}(\setminus A)$ designates the counterfactual operation where we remove all alternative causes from the scene. Once the alternatives are removed, we run the whether-cause test to see whether the candidate cause would have been enough to bring about the outcome on its own. In the situation where there are no alternative causes, the test reduces to a simple test of whether-causation.

We consider an object to be a potential alternative cause if there exists a counterfactual contingency where it would have been a whether-cause of the outcome. To test this, we simulate each counterfactual contingency, removing every subset of objects (excluding the target ball) and checking whether the candidate alternative was a whether-cause in any

²Note that unlike whether-causation, how-causation is binary. We only run a single simulation to determine whether the candidate cause is a how-cause.

of these contingencies.³ For example, in scenario 5, the blue ball is a whether-cause in the counterfactual contingency where ball A is removed. Thus, we consider the blue ball to be an alternative cause and remove it as part of the sufficient-cause test. If the blue ball hadn't been present, then ball A would have been a whether-cause, thereby satisfying our definition of sufficient-cause.

While this definition of sufficient-causation can help us explain why participants think ball A is a cause in scenario 5, it also makes the counter-intuitive prediction that the blue ball is a sufficient-cause in this situation. If we ran our test on the blue ball, we would remove the alternative cause (ball A) and then find that the blue ball is a whether-cause in this counterfactual contingency. This assessment is problematic: the blue ball clearly didn't cause ball B to go into the gate (Chang, 2009; Gerstenberg et al., 2021). To address this concern, we build on the definition of actual causation in Halpern and Pearl (2005). We further constrain sufficient-causation to check whether the events in the counterfactual contingency match the events that actually happened. The events we consider are the outcome, the obstacle collisions, and the button presses. We exclude events like balls entering the scene and balls colliding with the walls. Events are defined by the objects involved in them, and not the fine-grained details of their timing. Note that this check for whether events between situations match is asymmetric: events in the counterfactual contingency need to match those in the actual situation, but not vice versa.

To illustrate how this constraint impacts the sufficient-cause test consider scenario 5 again. Here ball A collides with ball B in the actual situation and also in the counterfactual contingency where we remove the blue ball. Ball A is also a whether-cause in this counterfactual contingency so we determine that it is a sufficient-cause. However, when we run the test for the blue ball, the story is different. In the counterfactual contingency where we remove ball A, the blue ball collides with ball B. This event does not happen in the actual situation. Even though the blue ball is a whether-cause in this contingency, the events that make it so didn't take place in the actual situation, so it doesn't satisfy our definition of sufficiency.

Sample Cases. We illustrate the model computations for each component of the model using the first four cases from Figure 4. Table 1a shows the computed aspect values for these cases. As we've noted above, in scenario 1 ball A is a whether-cause and a how-cause. It is also a sufficient-cause because there are no alternative causes. In scenario 2, ball A is a whether-cause but not a how-cause. It is also a sufficient-cause due to the fact that again there are no alternatives. In scenario 3, ball A is neither a whether-cause nor a sufficient-cause, but it is a how-cause. And in scenario 4, ball A isn't a whether-cause, a how-cause, or a sufficient-cause.

Semantics Module

We define a semantics that maps from people's causal representation of what happened to causal expressions. We consider four causal expressions: "affected", "enabled", "caused", and "made no difference". In defining the semantics for each of the expressions,

³When checking whether an object is a potential alternative cause, we ran the test for whether-causation with a single deterministic simulation. Running multiple simulations across every contingency proved to be prohibitively expensive. In principle, one could run multiple simulations for each contingency to capture uncertainty about which objects qualify as potential alternative causes.

Table 1

Model predictions for scenarios 1–4 shown in Figure 4. a) Aspect values computed for each of the different scenarios. b) Semantic valuations for each of the different causal expressions in those scenarios on the basis of the aspect values. c) Literal listener distributions over scenarios given a particular utterance. These are computed by normalizing the semantic values across scenarios. d) Speaker distributions for a first-level pragmatic speaker. These are computed by renormalizing the distribution of the literal listener across utterances.

a) Aspect Values					b) Semantic Values				
Scenario	1	2	3	4	Scenario	1	2	3	4
Whether	1.00	1.00	0.00	0.00	No Difference	0.00	0.00	0.20	1.00
How	1.00	0.00	1.00	0.00	Affected	1.00	1.00	1.00	0.00
Sufficient	1.00	1.00	0.00	0.00	Enabled	1.00	1.00	0.00	0.00
					Caused	1.00	0.00	0.00	0.00

c) Literal Listener Distributions					d) Speaker Distributions				
Scenario	1	2	3	4	Scenario	1	2	3	4
No Difference	0.00	0.00	0.17	0.83	No Difference	0.00	0.00	0.33	1.00
Affected	0.33	0.33	0.33	0.00	Affected	0.18	0.40	0.67	0.00
Enabled	0.50	0.50	0.00	0.00	Enabled	0.27	0.60	0.00	0.00
Caused	1.00	0.00	0.00	0.00	Caused	0.55	0.00	0.00	0.00

we will use \mathcal{W} , \mathcal{H} , and \mathcal{S} as a shorthand for $\mathcal{W}(A \rightarrow e)$, $\mathcal{H}(A \rightarrow \Delta e)$, and $\mathcal{S}(A \rightarrow e)$, respectively.

“Affected”. We define “affected” as

$$\text{AFFECTED}(A \rightarrow e) = \mathcal{W} \vee \mathcal{H} \vee \mathcal{S}. \quad (4)$$

A affected the outcome e if A was a whether-cause, a how-cause, a sufficient-cause, or any combination of the three. “Affected” is the most inclusive causal expression. If the candidate shows any of the different aspects of causation then it affected the outcome. According to this definition, ball A affected ball B’s going through the gate in all of the sample scenarios in Figure 4 except scenario 4.

“Enabled”. We define “enabled” as

$$\text{ENABLED}(A \rightarrow e) = \mathcal{W} \vee \mathcal{S}. \quad (5)$$

For A to have enabled e it must have either been a whether-cause, a sufficient-cause, or both. For example, in scenario 2 of Figure 4, ball A enabled ball 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. Scenario 6 demonstrates that including sufficient-causation is important for this causal expression. 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 there is a strong intuition that ball A enabled ball B’s going through the gate.

“Caused”. We define “caused” as

$$\text{CAUSED}(A \rightarrow e) = \mathcal{H} \wedge (\mathcal{W} \vee \mathcal{S}) \wedge_{\sigma} \mathcal{M} \wedge_{\sigma} \mathcal{U}. \quad (6)$$

A caused e when it was a how-cause of the outcome, and either a whether-cause or sufficient-cause (or both). In addition to these counterfactual components of the semantics, we further require that the candidate cause A was initially moving \mathcal{M} rather than stationary. This requirement is a soft condition as indicated by \wedge_{σ} . This means that, even if ball A was stationary at the beginning, it can still be said to have caused the outcome, but the probability of doing so is less than if A was moving. The degree of softening is controlled by a parameter σ . Prior work suggests that movement affects people’s causal intuitions (Mayrhofer & Waldmann, 2016; White, 2014), and we manipulate the movement of the candidate cause in our stimuli. For example, scenario 7 and scenario 8 are identical in terms of their causal aspects, but in scenario 7, ball A is moving while in scenario 8 it is stationary. As we will see below, participants are more inclined to choose “caused” in scenario 7 than in scenario 8.

Additionally, our definition for “caused” (softly) requires that the candidate cause A be unique \mathcal{U} . This requirement is motivated by the observation that the expression “ A caused B ” is ambiguous between two senses of “caused”. On the one hand, it could be taken to mean that A was *a cause* of the outcome, on the other, it could mean that A was *the cause* of the outcome. Prior work suggests that when multiple causes play a role in an outcome’s occurrence, people are inclined to distribute the causal contribution between them (Gerstenberg et al., 2021; Lagnado, Gerstenberg, & Zultan, 2013; Langenhoff, Wiegmann, Halpern, Tenenbaum, & Gerstenberg, 2021; White, 2014; Wu, Sridhar, & Gerstenberg, 2023). This issue becomes salient in cases like scenario 9. Here 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 gate”, doing so elides the equally important role that the blue ball played. It seems more appropriate to say that “ A caused B ” if A was a unique cause.⁴ Like \mathcal{M} , \mathcal{U} is a soft requirement. The degree of softening is also controlled by σ , the same parameter that determines the level of softening for the movement feature. We defined uniqueness in the following way: A contacted B, and no other candidate cause contacted B.

“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 by the softer constraints that a cause must be moving and that it be unique.

“Made no difference”. The expression “made no difference” asserts a lack of causal connection between the candidate cause and the outcome. We define the expression as the conjunctive negation of each of our causal aspects.

$$\text{NO DIFFERENCE}(A \rightarrow e) = \neg \mathcal{W} \wedge \neg_{\nu} \mathcal{H} \wedge \neg \mathcal{S}. \quad (7)$$

In words, A made no difference to e if it is not a whether-cause of e , not a how-cause of e , and not a sufficient-cause of e . The requirement that A not be a how-cause is soft (which

⁴This same competition does not seem to apply to the other expressions. If we say that “Ball A affected ball B’s going through the gate” it’s perfectly fine for the blue ball to have affected ball B as well. The same seems true for “enabled”.

we represent with the soft-not \neg_ν). This softening is intended to capture an ambiguity in the meaning of “made no difference” that we noted in cases like scenario 3. Here, ball B is headed toward the gate on its own when ball A comes up behind it and pushes it along. Ball A is a how-cause because it affected the fine-grained process, but it is neither a whether-cause nor a sufficient-cause. Even though ball A is a how-cause, it still seems reasonable here to say that ball A “made no difference” to ball B’s going through the gate because ball A made no difference to *whether* ball B went through the gate (it only made a difference to *how* it went through). In our experiment, we fit a parameter ν to capture the probability of responding “made no difference” even when the candidate was a how-cause.

Sample Cases. Table 1b shows how the semantic evaluations of our sample scenarios are shaped by their underlying aspect evaluations. In scenario 1, ball A is a whether-cause, a how-cause, and a sufficient-cause, so it satisfies the definition for “affected”, “enabled”, and “caused”, but not “made no difference”. In scenario 2, ball A is a whether-cause and a sufficient-cause. It satisfies the definition for “affected” and “enabled”, but because it is not a how-cause, it does not satisfy the definition for “caused”. In scenario 3, ball A is a how-cause so it satisfies the definition of “affected” but not “enabled” or “caused”. It also weakly satisfies the definition for “made no difference” due to the softening parameter that allows this expression to be applied even when the candidate is a how-cause. The softening parameter in this case was set to 0.2. Finally in scenario 4, ball A does not pass any of the counterfactual tests. As such, the only causal expression it satisfies is “made no difference”.

Pragmatics Module

Our semantics links the underlying psychological representation to the meanings of the four different causal expressions. While our semantics tells us when each of these expressions is true,⁵ it doesn’t tell us how people choose a particular expression in context. This poses a difficulty when multiple expressions are true. When that’s the case, prior work has shown that people choose their utterances on the basis of a variety of pragmatic conversational principles (Grice, 1989).

The rational speech acts (RSA) model quantitatively captures pragmatic reasoning in language use (Frank & Goodman, 2012; Goodman & Frank, 2016). The model posits a speaker and a listener who reason recursively about each other’s mental states in order to choose utterances that are both true and informative. Our implementation of the RSA model starts with a literal listener L_0 . Given an utterance with a causal expression, the literal listener infers a distribution over scenarios where any scenario that is consistent with the utterance is equally probable. Formally,

$$P_{L_0}(s|u) \propto \mathcal{M}(s, u). \quad (8)$$

The meaning function \mathcal{M} is our semantics. It takes a scenario s and an utterance u , and returns a semantic value $\in [0, 1]$. P_{L_0} is a distribution on scenarios where the semantic values are normalized to sum to one within each utterance. The change in values from Table 1b to Table 1c illustrates how the literal listener transforms the semantic values into

⁵Technically, the causal expressions themselves do not have truth values. Rather utterances containing the causal expressions have truth values. For the purposes of exposition, we will refer to the truth of causal expressions.

a distribution over scenarios. In this example, we limit the set of scenarios over which the listener reasons to scenarios 1–4 from Figure 4. Later, we apply the model to the full set of scenarios in our experiment.

Next, we define a level-1 pragmatic speaker who chooses an utterance so that the literal listener is likely to infer the scenario that the speaker observed. Given a scenario s , the probability that a speaker will choose an utterance u is proportional to the literal listener’s assessment of the probability of s given u . Formally,

$$P_{S1}(u|s) \propto P_{L0}(s|u)^\lambda. \quad (9)$$

Table 1c and 1d show that this step simply amounts to another round of re-normalization. Whereas the literal listener normalized the semantic values across scenarios (the columns of Table 1), the pragmatic speaker now normalizes the literal listener probabilities across the utterances (the rows of Table 1). Examining the speaker distribution in Table 1d shows an interesting consequence: in scenario 1, the pragmatic speaker favors the expression “caused” over the other two alternatives even though all of them are equally true (see the semantics in Table 1b). This is because “caused” is the most informative expression. It is semantically more restrictive and thus true of a smaller set of scenarios. The hierarchical re-normalization of RSA results in the speaker favoring this more informative utterance, capturing people’s tendency to be informative.

The λ parameter in the formula above is a softmax parameter which controls the optimality of the informative inference. By increasing this parameter, the effect of informativity intensifies. As λ increases, the mass of the distribution concentrates around the most probable response. As λ approaches zero, the distribution becomes increasingly flat, becoming uniform at $\lambda = 0$. The example in Table 1 assumes $\lambda = 1$.

Recursive reasoning in RSA can be repeated to an arbitrary depth. We can construct a pragmatic listener that reasons about a level-1 pragmatic speaker, and a level-2 pragmatic speaker that reasons about a level-1 listener. Additional levels of recursion increase the effect of informativity, but also impose increased computational costs. In this work we model participant speakers with a level-2 pragmatic speaker, and participant listeners with a level-1 pragmatic listener.

Sample Cases. Tables 1c and d illustrate pragmatic inferences for a listener and a speaker. For each utterance, the literal listener (1c) normalizes the semantic values across the set of scenarios. If the utterance is true in multiple scenarios, the literal listener assigns equal probability to each of those scenarios. For example, if the literal listener hears that “Ball A enabled ball B to go through the gate”, the literal listener will infer that scenarios 1 and 2 are equally possible because the utterance truthfully describes those scenarios. The literal listener’s inference can be modulated by the graded semantic evaluation. For the utterance “Ball A made no difference to ball B’s going through the gate”, the literal listener assigns most probability to scenario 4, but also applies some probability to scenario 3 because that scenario weakly satisfies the definition for “made no difference” due to the softening parameter.

Table 1d illustrates the inference for the pragmatic speaker. The pragmatic speaker normalizes the probabilities of the literal listener across utterances, resulting in a relative increase in the probability of utterances that are more informative (i.e. true of fewer scenarios). Because the utterance “Ball A caused ball B to go through the gate” is only true

in scenario 1, the pragmatic speaker assigns this utterance more probability than the corresponding utterances with “enabled” and “affected”. This reflects a pragmatic norm to choose the most informative utterance that is true in a given scenario.

Experiment 1: Validating Causal Expression Semantics

According to our model semantics, the three causal expressions “caused”, “enabled”, and “affected” overlap in meaning. However, some of the expressions are more specific than others. They refer to a smaller set of possible situations and therefore are more informative (see Figure 3, semantics module). The most specific and informative expression is “caused”, followed by “enabled”, and then “affected”. In order for “caused” to be true, the conditions for “enabled” and “affected” must also be satisfied and so we say that “caused” implies “enabled” and “affected”. Similarly, in order for “enabled” to be true, the conditions for “affected” must also be satisfied, so under our semantics, “enabled” implies “affected”.

This scale of implication further grounds our model pragmatics, which predicts that participants will favor more informative expressions when multiple expressions are true of a given situation. Moreover, our model predicts that in a situation in which a listener hears a less specific utterance such as “enabled”, they will infer that the more specific utterance, “caused”, is not true. This is an example of a scalar implicature (Hirschberg, 1985). Scalar implicatures arise when a set of words exist on a scale of specificity, and using a less specific word implicates (but does not imply) that the more specific words on the scale are not true. A classic example of a scalar implicature is the relationship between “all” and “some” (Degen, 2015; Goodman & Stuhlmüller, 2013). “All” and “some” form a scale of specificity. “All” is more specific than “some”. If someone tells you that “some of the cups have soda in them”, you will likely infer that not all of the cups have soda in them, because if they did, the speaker would have told you so. This inference from the use of the weaker expression to the conclusion that the stronger expression is not true is the scalar implicature.

We propose that a similar relationship holds among the causal expressions “caused”, “enabled”, and “affected”. To test this assumption, we examined participants’ intuitions about which expressions imply which others, and participants’ willingness to endorse statements consistent with our model pragmatics. Experiment 1 includes three studies: an initial study where we tested the baseline acceptability of our causal expressions in a variety of sentence frames, a second study to assess whether participants’ intuitions about the causal expressions are consistent with our semantics, and a third study to assess whether participants cancel implicatures as predicted by our pragmatic account.

Experiment 1A: Norming study

If saying that “The new technology caused the change.” implies that it enabled and affected it, it must at a minimum be acceptable to use all three of these verbs in the same sentence frame. To test our hypotheses, we need a collection of sentence frames where all three causal expressions are acceptable. To collect these stimuli, we first ran a norming study that examined the acceptability of using the different causal expressions in a set of sentence frames. The sentence frames that participants rated acceptable across all the expressions in this norming study then served as the base for the stimuli in our follow-up experiments where we tested our primary hypotheses.

Table 2

Sample sentence frames from the norming study. The top ten sentence frames had median ratings above the midpoint of the scale for all three causal expressions. These sentence frames were used to construct stimuli in the follow-up experiments. The bottom three are a sample of the sentence frames that were excluded. For these sentence frames, the median response for at least one expression was at the midpoint of the scale or below.

Included Sentence Frames
1. The dry weather ____ the wild fire.
2. The CEO's decision ____ the outcome.
3. The new technology ____ the change.
4. The Sacklers' greed ____ the opioid epidemic.
5. The sunny weather ____ the tree's growth.
6. More stipends ____ the increase in student admissions.
7. The sun ____ the drying of the clothes.
8. Metastasis ____ cell growth.
9. Diversification ____ new monetary policies.
10. The algae buildup in the ocean ____ the migration of certain species of fish.

Examples of Excluded Sentence Frames
1. The collapse of Lehman Brothers ____ the financial crisis.
2. Janelle's working hard ____ her success.
3. Turning off the life support ____ the patient's death.

Methods

All experiments were approved by Stanford's IRB (#48665). Experiment 1A was developed and deployed using the jsPsych experiment library (De Leeuw, 2015). We pre-registered our data-collection paradigm and analysis plans on the Open Science Framework: <https://osf.io/kx5fg>. The data, study materials, and analysis code for all experiments in this paper are available here: https://github.com/cic1-stanford/causal_language

Participants. We recruited 51 participants (*age*: M = 35, SD = 12, *gender*: 25 female, 23 male, 3 non-binary, *race*: 37 White, 6 Asian, 6 Black/African-American, 2 other) via Prolific. We restricted selection to participants who had completed at least 10 previous experiments and have an overall 95% experiment approval rating. All participants were fluent in English and based in the United States (we use the same inclusion criteria for the subsequent studies in Experiment 1). We excluded one participant who failed to pass the attention check, leaving a total of 50 participants for analysis. Participants were paid \$2 for 10 minutes of work.

Stimuli. We developed a set of 20 sentence frames to test for acceptability with each of our causal expressions. For each sentence frame, we created three items by substituting in each of the three causal expressions. In total, we had 60 stimuli sentences. The full set of stimuli included a variety of scenarios across many different domains of causation including medicine, weather, and finance. Table 2 shows some examples, and Appendix A shows the full set of sentence frames. We also included six attention checks, two for each causal expression, which were designed to be either obviously acceptable or obviously unacceptable.

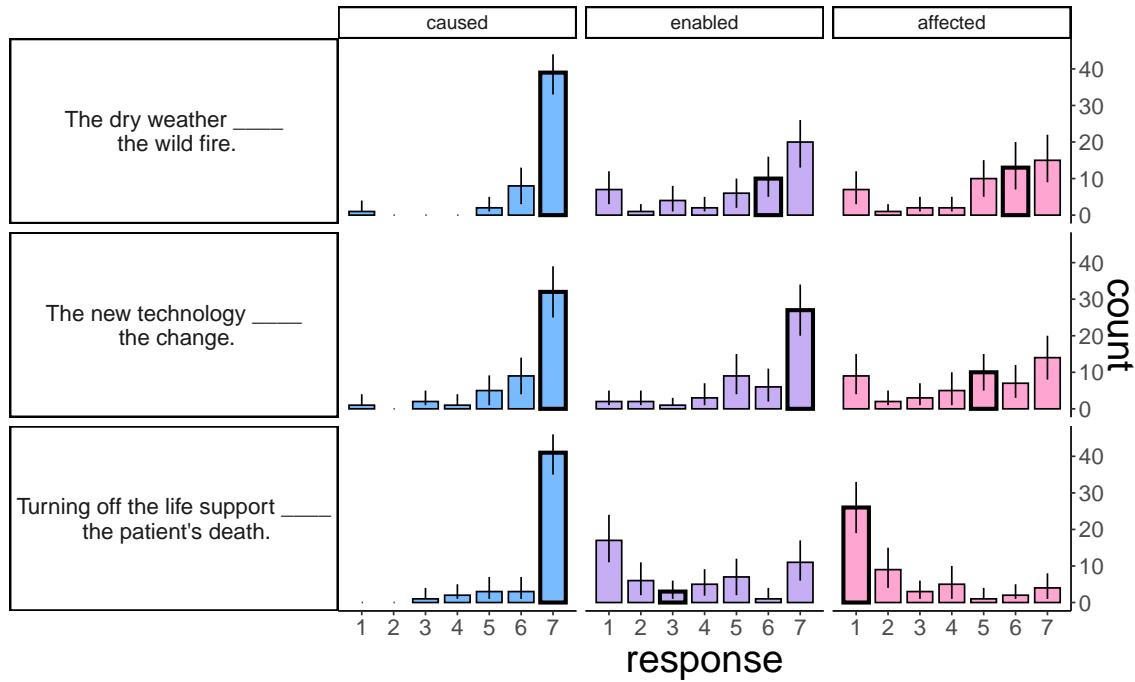


Figure 5. Experiment 1A. Sample sentence frames considered for inclusion in our experiment with histograms for each causal expression showing participant ratings of acceptability. Median responses are indicated by a bold outline. The top two sentences were included because participant median ratings on all three verbs were above the midpoint of the scale. The bottom sentence was excluded because participant ratings for both “enabled” and “affected” fell below the threshold. Error bars represent 95% bootstrapped confidence intervals.

Participants who failed to answer on the correct side of the scale for more than two attention checks were excluded from analysis.

Procedure. We instructed participants that they would see a series of 66 sentences and rate their acceptability. We provided minimal guidance to the meaning of acceptability encouraging participants to trust their own intuitions of what sounds “natural”. We provided participants with an example that we labeled as acceptable (“Working long hours caused Pat to feel tired.”) and an example that we labeled unacceptable (“Working long hours affected Pat to feel tired.”). After reading these instructions, participants proceeded to the main task. Items were presented one by one in randomized order and led by the prompt “Is this an acceptable English sentence?”. Participants provided ratings on a 7-point Likert scale with the endpoints labeled “definitely no” and “definitely yes”, and the midpoint labeled “unsure”. Participants had to provide a judgment on each item to continue.

Results and Discussion

Our aim was to collect a set of 10 sentence frames for which all three causal expressions were acceptable. We defined acceptability as a median rating of 4 (the scale midpoint) or higher. 15 of 20 frames met this threshold. In our exploratory analysis, we found that 10

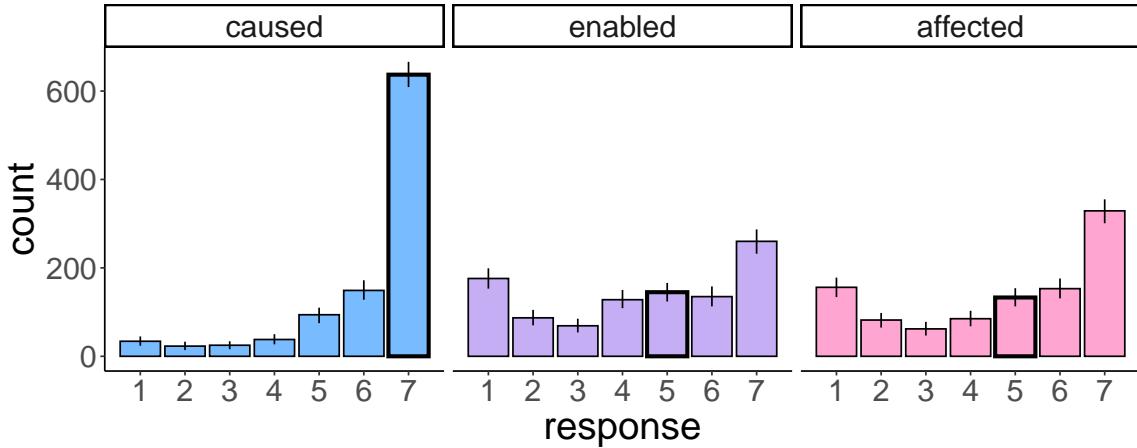


Figure 6. Experiment 1A. Overall responses for each verb aggregated across items. In general, sentence frames with “caused” were rated highly acceptable. Sentence frames with “enabled” and “affected” showed more variance in the distribution of responses. *Note:* Median responses are indicated by a bold outline. Error bars represent 95% bootstrapped confidence intervals.

sentence frames had a median acceptability of 5 or higher for all three causal expressions. We selected these 10 sentence frames as our stimuli for our follow-up experiments. Figure 5 shows histograms of acceptability ratings for three example sentence frames. In the top two frames, participants found all three expressions acceptable. In the bottom sentence frame, only “caused” was rated acceptable. Therefore, we didn’t include this sentence frame in subsequent experiments.

In general participants found items with “caused” more acceptable than items with “enabled” or “affected” for our sentence frames. Figure 6 shows histograms of participant responses for each verb, aggregated across items. The histogram for “caused” skews strongly toward the acceptable end of the scale. On the other hand the histograms for “enabled” and “affected” show more variance, though still tend overall toward acceptability. Considering individual frames, the median rating for “caused” was above the midpoint of the scale on all frames. The rating for “enabled” was above the midpoint in 12 out of 20 frames. The median rating for “affected” was above the midpoint in 14 out of 20 frames.

Experiment 1B: Semantic Relations of Causal Expressions

The stimuli collected in Experiment 1A provide us with a set of sentences to test the semantics of our causal expressions. We want to know whether people’s intuitions about the relationships among the expressions align with the structure of the semantics module as shown in Figure 3. When people say that “A caused B”, does that imply that “A enabled B” and “A affected B”? Similarly when people say that “A enabled B”, does that imply that “A affected B”? While these implications from stronger to weaker should hold if our semantics is true, the reverse is not the case. It should be acceptable for A to enable or affect B without causing it. The implication is uni-directional.

To test this prediction, we augmented our stimuli from the previous experiment with

an additional “but it didn’t ____ it” clause. For example, the sentence frame “The sunny weather ____ the tree’s growth.” became “The sunny weather ____ the tree’s growth, but it didn’t ____ it.”⁶ To create our stimuli, we substituted each pair of expressions into the blanks in these sentence frames. When the order of the expressions goes from a more specific verb to a less specific verb (e.g. “The sunny weather caused the tree’s growth, but it didn’t affect it.”), we hypothesized that participants would find the sentence unacceptable. This is because, under our semantics, sentences like these express a contradiction. If “caused” implies “affected”, then saying that “A caused B, but it didn’t affect B” implies that “A affected B, but it didn’t affect B”. On the other hand, we hypothesized that the reverse ordering of the expressions, where the less specific expression comes first followed by the more specific expression, should be acceptable (e.g. “The sunny weather enabled the tree’s growth, but it didn’t cause it.”). According to our semantics there are possible scenarios where the more specific expression (e.g. “caused”) is not true while the less specific expression is (e.g. “enabled”).

Methods

Experiment 1B was developed and deployed using the jsPsych experiment library (De Leeuw, 2015). We pre-registered our data-collection paradigm and analysis plans on the Open Science Framework: <https://osf.io/2un9v>.

Participants. We recruited 55 participants (*age*: M = 42, SD = 14, *gender*: 31 female, 24 male, *race*: 46 White, 4 Asian, 3 Black/African American, 2 other) online using the Prolific platform. We excluded 2 participants that failed to pass the attention check. Participants were paid \$2 for 10 minutes of work.

Stimuli. We took all ten of the sentence frames collected in our preliminary study and augmented them with the “but it didn’t ____ it” clause as described above. To create our experiment items, we permuted each pair of causal expressions in each frame, leading to a total of 60 items. Table 3 shows six sample items for this experiment on the top of the table. In addition to our primary experimental items, we included the same six attention checks as we had in the preliminary experiment. As in the previous experiment, participants who failed more than two of these attention checks were excluded from analysis.

Procedure. The procedure for this experiment was very similar to that of Experiment 1A. We instructed participants that their task was to rate the acceptability of 66 sentences. Again, we informed participants that we were interested in their intuitions of whether or not each sentence seemed “natural”. We provided two example sentences with the same form as our experimental items. In one sentence a more specific verb followed a less specific verb, and in the other sentence the reverse was true. Unlike in the preliminary experiment, we did not indicate whether either of these example sentences was acceptable or not. Participants proceeded to provide ratings on the same Likert scale as in Experiment 1A. Again, the item order was randomized.

Hypotheses. We hypothesized that items where the less specific verb preceded the more specific verb (affected → caused, affected → enabled, enabled → caused) would be more acceptable overall than the corresponding items with the reversed orders (caused →

⁶Shibatani (1976) uses a similar construction to illustrate the more restrictive meaning of lexical causatives relative to the periphrastic causative “cause”.

Table 3

Experiment 1. Example items from the Semantics Experiment (Experiment 1B) and the Pragmatics Experiment (Experiment 1C). In the Semantics Experiment, items with verb orderings that are contradictory under our semantics are presented on the left, and corresponding items that are consistent with our semantics are shown on the right. In the Pragmatics Experiment, items that use the “in fact” phrasing to cancel the implication are shown on the left, and corresponding items that use the “in fact” phrasing to cancel the implicature are shown on the right.

Semantics Experiment (Experiment 1B)	
Contradictory Ordering	Non-Contradictory Ordering
The Sackler’s greed caused the opioid epidemic, but it didn’t enable it.	The Sackler’s greed enabled the opioid epidemic, but it didn’t cause it.
The sunny weather caused the tree’s growth, but it didn’t affect it.	The sunny weather affected the tree’s growth, but it didn’t cause it.
Metastasis enabled cell growth, but it didn’t affect it.	Metastasis affected cell growth, but it didn’t enable it.
Pragmatics Experiment (Experiment 1C)	
Implication Cancellation	Implicature Cancellation
The dry weather caused the wild fire, in fact it enabled it.	The dry weather enabled the wild fire, in fact it caused it.
The CEO’s decision caused the outcome, in fact it affected it.	The CEO’s decision affected the outcome, in fact it caused it.
The new technology enabled the change, in fact it affected it.	The new technology affected the change, in fact it enabled it.

affected, enabled → affected, caused → enabled). We broke down this hypothesis into three sub-hypotheses, one for each pair of causal expressions. For example, we predicted that enabled → caused would be more acceptable than caused → enabled. To test these hypotheses, we conducted a Bayesian regression analysis. We fit a hierarchical Bayesian ordinal regression predicting participant responses from the verb ordering in the given item. Verb ordering was represented as a six-level factor, one level for each permutation of our causal expressions (enabled → affected, caused → enabled, etc.). Additionally, we included random intercepts for participants and sentence frames. We fit the regression using the `brms` package (Bürkner, 2017) in the programming language R (R Core Team, 2019). Default priors were assumed for all parameters.

We computed linear contrasts on the levels of the verb ordering factor of the fitted regression model to test our hypotheses. For each causal expression pairing, we subtracted the posterior estimate of the parameter on the ordering where the first verb implied the second from the posterior estimate for the reverse ordering. For example, we took the distribution representing the parameter estimate for the verb ordering caused → affected and subtracted it from the distribution representing the parameter estimate for the reverse

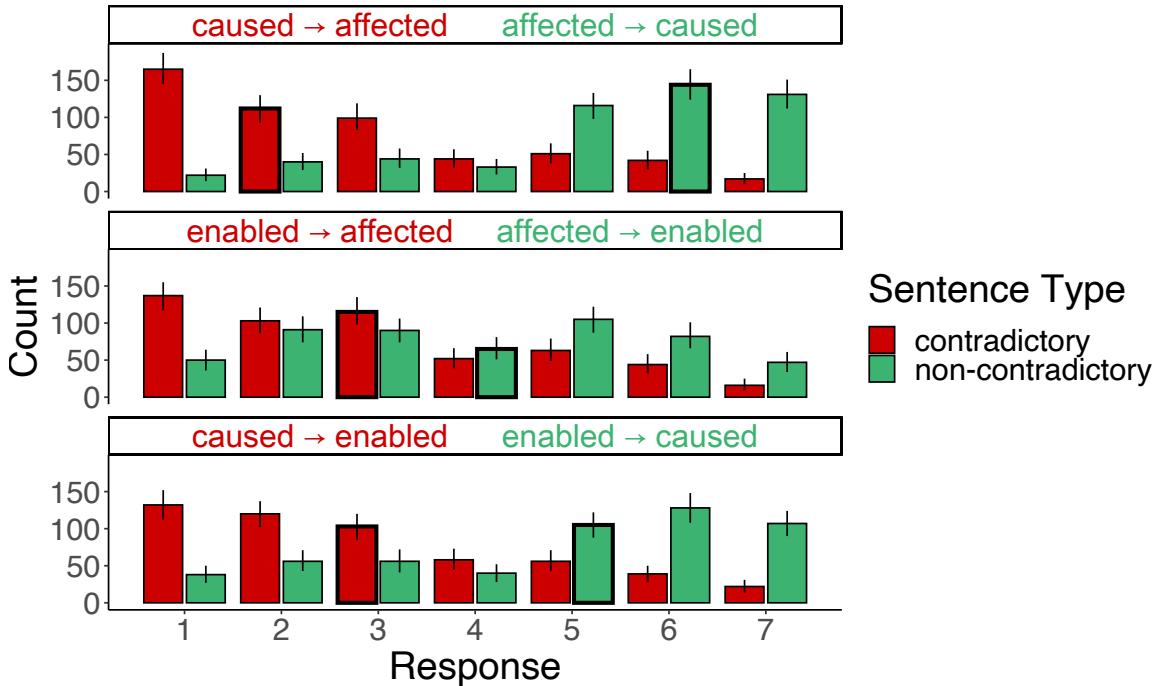


Figure 7. Experiment 1B. Histogram of the overall acceptability of contradictory items and non-contradictory items in the semantics experiment. Median ratings are indicated by bold bar outline. Across all pairings, contradictory items are overall less acceptable than non-contradictory items, and the median rating for the contradictory items is below the midpoint of the scale for all pairings of causal expressions. For the non-contradictory items, the median rating of “affected” → “caused” and “enabled” → “caused” is above the midpoint of the scale, while for “affected” → “enabled” the median is at the midpoint of the scale. Error bars reflect bootstrapped 95% confidence intervals.

ordering affected → caused. We considered each sub-hypothesis to be confirmed if 95% of the resulting contrast distribution was above zero.

Results

Confirmatory analyses. Figure 7 shows participant responses. Each row shows the results for a particular pair of expressions. The red bars represent responses on orderings that were contradictory under our semantics, and the green bars represent responses on orderings that were consistent with our semantics. In general, the green bars skew to the right of the red bars, indicating that participants found items which are consistent with our semantics overall more acceptable than items which are contradictory under our semantics. For example, participants rated items like “The dry weather caused the wild fire, but it didn’t enable it.” less acceptable than the reversal “The dry weather enabled the wild fire, but it didn’t cause it.”

Table 4 shows the results of our confirmatory hypothesis tests. All hypotheses were confirmed. 95% of the posterior density for each contrast distribution lies above zero, indicating there is a credible difference between the orderings for each verb pairing.

Exploratory analyses. Our confirmatory analyses tested the difference between the contradictory and the non-contradictory verb orderings. In addition to this relative effect, we were also interested in whether participants generally thought that the contradictory frames were “unacceptable” and the non-contradictory ones “acceptable”. We conducted an exploratory analysis where we looked at the median rating for each verb ordering. We considered an ordering to be rated overall acceptable if the median rating was above the midpoint of the scale (4), and unacceptable if it was below.

The bold outlines in Figure 7 highlight the median values for each verb ordering. For all three contradictory orderings, the median rating was below the midpoint of the scale. For two of the three non-contradictory orderings, the median rating was above the midpoint of the scale. The one exception was “affected” → “enabled”, for which the median rating was at the midpoint.

Discussion

The results of our analysis generally confirm our model semantics. Overall, participants rated sentences that implied a contradiction under our semantics as unacceptable and sentences that were consistent as acceptable. This suggests that, in line with the scale of specificity that we hypothesized, participants believe that “caused” implies “enabled”, and that “enabled” implies “affected”.

One notable exception is the finding that the median rating for sentences where “affected” preceded “enabled” was at the midpoint of acceptability. There was substantial variance in participants’ acceptability ratings for this ordering (see Figure B1). For some items, such as “The CEO’s decision affected the outcome, but it didn’t enable it.”, the distribution of responses skews toward acceptability. For other items, such as “The sunny weather affected the tree’s growth, but it didn’t enable it.”, the distribution skews toward non-acceptability. Still for others, such as “Metastasis affected cell growth, but it didn’t

Table 4

Summary of the confirmatory hypothesis test results for Experiment 1B. For each pairing of verbs, we computed the contrast distribution by subtracting the posterior distribution of the verb ordering that went from the stronger verb to the weaker verb from the corresponding distribution for the ordering from weaker verb to the stronger verb (e.g. subtracting the posterior for the caused → affected ordering from the posterior for the affected → caused ordering). The ‘Posterior Estimate’ column represents the mean of the resulting contrast distribution. The middle column reports the lower bound of the credible interval. 95% of the contrast distribution lies above this bound. The ‘Posterior Probability’ column reports the proportion of samples from the posterior density favoring the hypothesis that the contrast distribution lies above zero. For all pairings, all samples spoke in favor of the hypothesis indicating very strong evidential support.

Verb Pairing	Posterior Estimate	Credible Interval Lower Bound	Posterior Probability
AFFECTED-CAUSED	1.47	1.34	~ 1
ENABLED-CAUSED	1.09	0.97	~ 1
AFFECTED-ENABLED	0.62	0.50	~ 1

enable it.”, the distribution is flat reflecting the uncertainty that we see at the aggregate level. Even though participants were generally more uncertain about sentences that went from “affected” to “enabled”, they were confident that sentences with the reverse ordering were unacceptable (e.g. “The dry weather enabled the wild fire, but it didn’t affect it.”).

Experiment 1C: Pragmatics of Causal Expressions

Next, we examine whether people exhibit the pragmatic behaviors we would expect based on our model semantics. As we noted above, when semantic scales have this type of informative hierarchy, they give rise to scalar implicatures. If a speaker uses a weaker verb on the scale, a listener will generally infer that the speaker doesn’t think the stronger verb is true of the situation being described.

A standard method to test for whether a particular expression is an implicature is to see whether it can be *cancelled* (Grice, 1989; Mayol & Castroviejo, 2013). Returning to the example of “all” and “some”, if a speaker tells a listener “Some of the cups have soda in them.”, the listener will likely infer that “Not all of the cups have soda in them.”. However, the speaker can cancel this implicature in their statement by adding an additional clause: “Some of the cups have soda in them, in fact all of them do.” The “in fact” clause adds additional information that is consistent with, but stronger than what was said before (Matsumoto, 1997). In contrast, if one tries to cancel the implication from “all” to “some” the result is less natural: “All of the cups have soda in them, in fact some of them do.” Here, the “in fact” clause offers redundant information. Because stating that “all the cups have soda in them” implies that “some of the cups have soda in them”, the “in fact” clause is merely repeating something that was already communicated.

The bottom half of Table 3 illustrates the contrast between sentences that use the “in fact” locution to cancel an implication or an implicature. Under our semantics, the sentences that cancel implicatures should be judged acceptable, while the sentences that cancel implications should appear unnatural and redundant.

Methods

Experiment 1C was developed and deployed using the jsPsych experiment library (De Leeuw, 2015). We pre-registered our data-collection paradigm and analysis plans on the Open Science Framework: <https://osf.io/ak5yd>.

Participants. We recruited 54 participants (*age*: M = 34, SD = 13, *gender*: 27 female, 25 male, 2 Non-binary *race*: 37 White, 6 Asian, 5 Black/African American, 1 American Indian/Alaska Native, 5 other) on the Prolific platform. We excluded 3 participants who failed to pass an attention check, leaving us with a total of 51 participants in our analysis. Participants were paid \$2 for 10 minutes of work.

Stimuli. As in the previous experiment, we took the set of ten sentence frames that we collected in Experiment 1A and augmented them. This time, instead of adding a “but it didn’t ____ it” clause, we added an “in fact it ____ it” clause. For example, the preliminary frame “The CEO’s decision ____ the outcome.” became “The CEO’s decision ____ the outcome, in fact it ____ it.” To create our full set of items we again permuted each pair of causal expressions in each frame, leading to a total of 60 items. Table 3 shows six

sample items on the bottom of the table. As in the previous two experiments, we included the same attention check items, with the same criteria for exclusion.

Procedure. The procedure for this experiment was nearly identical to that of the previous study. Participants saw the same instructions except the two sample sentences were replaced by a pair of examples using the “in fact” locution. The only other difference was the items themselves, which were constructed as explained above.

Hypotheses

We hypothesized that sentences where the “in fact” clause cancelled an implicature (affected → caused, affected → enabled, enabled → caused) would be overall more acceptable than sentences where the same clause attempted to cancel an implication (caused → affected, enabled → affected, caused → enabled). For example, sentences like the ones on the right side of the bottom half of Table 3 would be more acceptable than sentences like the ones on the left. As in the previous experiment, we fit a hierarchical Bayesian ordinal regression predicting participant response from verb ordering, and then computed linear contrasts on the levels of the verb ordering to test our hypotheses. We subtracted posterior estimates for the three verb orderings where the stronger verb preceded the weaker one from the posterior estimates of the corresponding verb orderings where the weaker verb preceded the stronger one. We tested each pair of causal expressions individually and considered the hypothesis to be confirmed if 95% of the contrast distribution was above zero. Our ordinal regression included random intercepts for participants and sentence frames, and we assumed default priors for all parameters.

We also included a control in the regression for the raw acceptability of the different causal expressions in each frame. As we noted in the discussion of Experiment 1A, the three causal expressions varied in their acceptability with “caused” being the most acceptable. We hypothesized that independent from any consideration of the implicature, the switch in acceptability from the less acceptable response to the more acceptable one could drive participant responses and confound our effect. For example, in the item pair, “The new technology affected/caused the change, in fact it caused/affected it”, participants might overall rate the affected → caused ordering more acceptable than the caused → affected order merely because “The new technology caused the change.” is more acceptable than “The new technology affected the change.”, and the “in fact” locution corrects from the less acceptable one to the more acceptable one.

To control for this possibility we computed an acceptability difference predictor for each item. We took the median acceptability scores for both verbs in that item from the preliminary experiment, and subtracted the acceptability of the first verb from the second. We computed this score for each item and included it as another predictor alongside the verb-ordering itself.⁷

⁷We performed a similar pair of controls in Experiment 1B. We tested one control where the acceptability value was just the median acceptability of the first verb in the item, computed from acceptability judgments provided for that verb and that sentence frame in the norming study. We tested a second control where we computed an acceptability difference, but in this case we subtracted the median acceptability of the second verb from the first. In both cases, the controls did not impact the results of our hypothesis tests. The posterior estimates for all hypotheses were positive and the 95% credible intervals excluded zero.

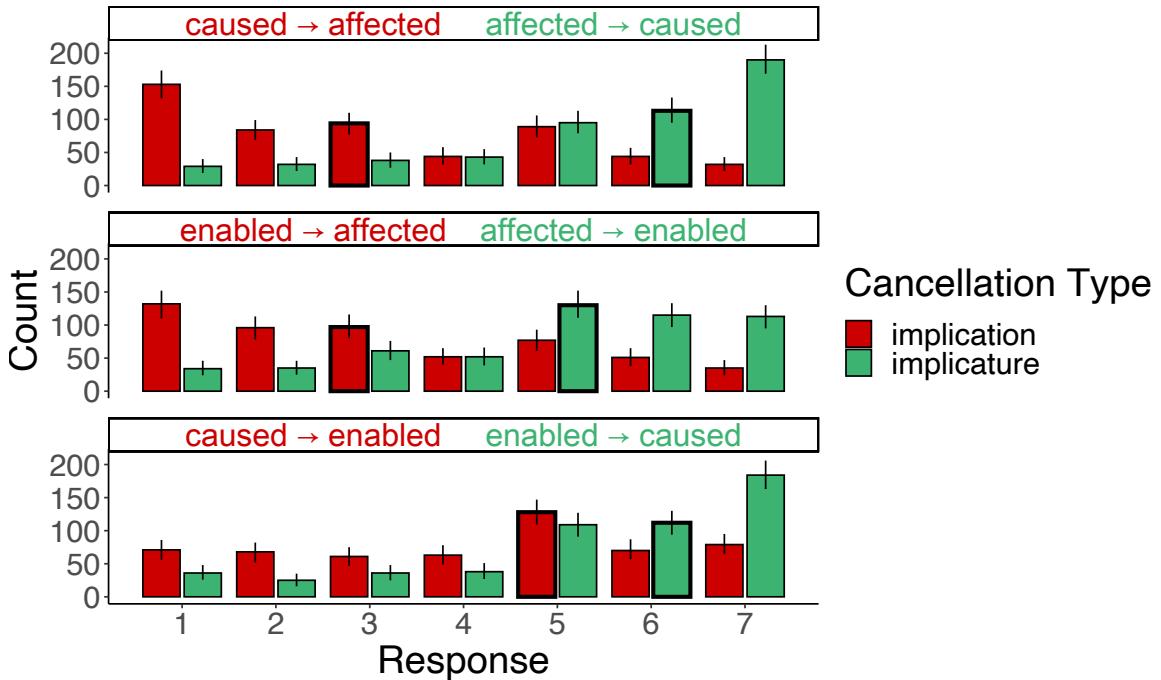


Figure 8. Experiment 1C. Histogram of overall acceptability of implicature cancellations vs. implication cancellations for each causal expression pairing. Median ratings are indicated by bold bar outline. Across all pairings, the implicature cancellations are more acceptable than implication cancellations, and the median rating for the implicature cancellations is always above the midpoint of the scale. For the implication cancellations, the median rating of “enabled” → “affected” and “caused” → “affected” is below the midpoint of the scale, while for “caused” → “enabled” the median is above the midpoint of the scale.

Results

Confirmatory analyses. Figure 8 shows participant responses. Again, each row represents a particular verb pair. Here, the red bars represent items that cancelled an implication, while the green bars represent items that cancelled an implicature. Again, we see that the green bars consistently skew to the right of the red bars. Overall, participants rated items that cancelled implicatures more acceptable than items that cancelled implications. For example, participants were more likely to give high ratings to items like “The sunny weather affected the tree’s growth, in fact it caused it.” than items like “The sunny weather caused the tree’s growth, in fact it affected it.”

Table 5 displays the results of our confirmatory hypothesis tests. Again, we see that 95% of each contrast distribution lies above zero, indicating there is a credible difference between each verb order in line with our pre-registered hypotheses.

Exploratory analyses. As in the previous study, we also assess whether participants judge that items with a particular verb ordering are acceptable regardless of the comparison with the reverse order. We computed the median acceptability for each verb ordering, aggregating across all items. The bolded outlines in Figure 8 present the results of our exploratory analysis. For all the verb orderings where a less informative verb pre-

Table 5

Summary of the confirmatory hypothesis test results for Experiment 1C. For each pairing of verbs, we computed the contrast distribution by subtracting the posterior estimate of the verb ordering that went from the stronger verb to the weaker verb from the corresponding estimate for the ordering from weaker verb to the stronger verb (e.g. subtracting the posterior for the caused → affected ordering from the posterior for the affected → caused ordering). The ‘Posterior Estimate’ column represents the mean of the resulting contrast distribution. The middle column reports the lower bound of the credible interval. 95% of the contrast distribution lies above this bound. The ‘Posterior Probability’ column reports the proportion of samples from the posterior density favoring the hypothesis that the contrast distribution lies above zero. For all pairings, all samples spoke in favor of the hypothesis indicating very strong evidential support.

Verb Pairing	Posterior Estimate	95% of posterior density above	Posterior Probability
AFFECTED-CAUSED	1.37	1.21	~ 1
AFFECTED-ENABLED	0.96	0.83	~ 1
ENABLED-CAUSED	0.73	0.57	~ 1

ceded a more informative verb (e.g. “The Sacklers’ greed affected the opioid epidemic, in fact it caused it.”) the median rating was above the midpoint of the scale, indicating that participants in general found implicature cancellations acceptable. In contrast, for the verb orderings where the more informative verb preceded the less informative verb (e.g. “The Sacklers’ greed caused the opioid epidemic, in fact it affected it.”), two were below the midpoint of the scale (enabled → affected and caused → affected) and one was above the midpoint (caused → enabled).

Discussion

The results of this experiment support the pragmatic assumptions of our model. Participants are more inclined to accept an “in fact” cancellation statement for verb orderings that give rise to scalar implicatures than for verb orderings where the first verb implies the second verb. For example, participants generally find statements such as “The new technology affected the change, in fact it caused it.” more acceptable than “The new technology caused the change, in fact it affected it.” Moreover, participants’ judgments of “in fact” statements that cancel implicatures generally skew to the acceptable side of the scale. The median acceptability is above the midpoint for all three of these orderings. The broad acceptability of the cancellation suggests that the use of the weaker verbs does indeed implicate that the stronger verbs aren’t true.

Notably, the acceptability of statements that cancel implications is more mixed. The median acceptability for all three of these orderings hovers around the midpoint of the scale. This contrasts with the unacceptable statements in Experiment 1B (red bars in Figure 7) where participants’ responses strongly skewed toward the unacceptable end of the scale. One way to make sense of this difference, is that the two unacceptable sentence types (e.g. “The new technology caused the change, but it didn’t affect it.” versus “The new technology caused the change, in fact it affected it”) are unacceptable for different reasons. In the “but it didn’t” constructions, the sentences that have verb orders where the

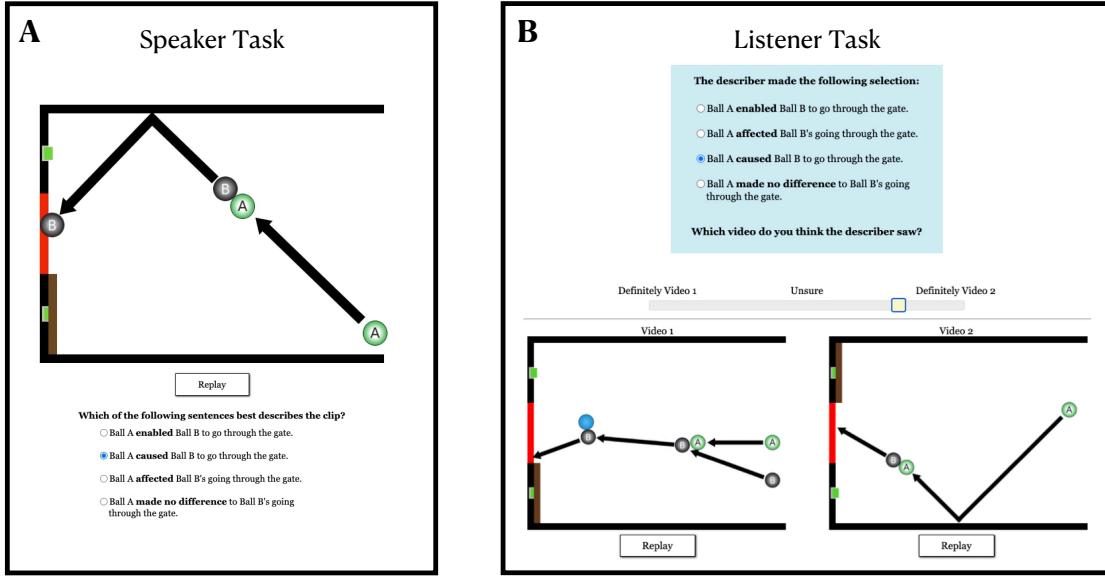


Figure 9. Sample trials from **A** the speaker task (Experiment 2), and **B** the listener task (Experiment 3). In the speaker task, participants chose one out of four utterances that best described what happened in the video clip. In the listener task, participants saw which one of the four utterances had been selected by a hypothetical speaker. Participants rated on a slider which one of two videos they thought the speaker had seen.

stronger verb comes first are contradictory. It is not possible for something to “cause” but not “affect” an outcome, and so these types of sentences are seen as highly unacceptable. On the other hand, for the “in fact it” constructions, following the stronger verb with the weaker verb isn’t contradictory, it is merely redundant. Stating the weaker verb in the second clause of the “in fact it” construction repeats something that was already implied by the first clause. A listener might find this strange, but not necessarily as unacceptable as an outright contradiction.

Interestingly, there was a lot of variance in participants’ responses for the “caused” → “enabled” ordering. For items such as “The sunny weather caused the tree’s growth, in fact it enabled it.” many participants felt that this sentence was acceptable. However, across frames with this ordering there is interesting variation in the distribution of responses. For some items, the responses skew more to the acceptable side of the scale, while for others the distribution of responses is more uniform across the scale. See Figure C1 for visualizations of all items with this verb order.

Experiment 2: Speakers choosing what to say

Having validated the linguistic principles underlying our model, we now turn to quantitative assessments against participant behavior. We investigate how people choose what causal expression best describes what happened in dynamic physical scenarios. In our experiment, participants viewed physical scenarios like those illustrated in Figure 2, and chose

from four causal expressions (“caused”, “enabled”, “affected”, and “made no difference”) the one that best describes the scenario. These physical scenarios allow us to quantitatively manipulate the different aspects of causation and see how this affects participants’ use of the different causal expressions.

Methods

Participants. We recruited 64 participants (*age*: M = 35, SD = 8, *gender*: 19 female, 43 male, 2 no response, *race*: 49 White, 6 Asian, 3 Black, 2 mixed race, 4 unclear or no response) online via Mechanical Turk using Psiturk (Gureckis et al., 2016). We excluded two participants from analyses who failed to select “made no difference” on an attention check video in which ball A lay still in a corner and thus clearly made no difference to the outcome. The experiment took 25 minutes on average (SD = 11), and participants were paid \$3.67.

Stimuli. We created 30 videos depicting physical scenarios, including the examples in Figure 4. All of the scenarios featured a ball labeled A, a ball labeled B, a red gate, a brown door, and two buttons controlling the door’s movement (see Figure 2). Some of the scenarios also included a blue ball or a brown box. We designed the scenarios to exhibit a range of values for each of the different causal aspects, thereby capturing a wide range of causal interactions. Schematics for all scenarios and corresponding aspect values are provided in Appendix D.

Procedure. We screened for potential bots by asking a simple natural language question. Participants then received instructions about the task. We introduced the domain and the different objects in it, and had participants watch a video illustrating a scenario with all of the objects from the domain. Participants were told that they would view scenes like this one and then be asked to choose one out of the four descriptions that best captured the scene they viewed:

1. “Ball A **caused** ball B to go through the gate.”
2. “Ball A **enabled** ball B to go through the gate.”
3. “Ball A **affected** ball B’s going through the gate.”
4. “Ball A **made no difference** to ball B’s going through the gate.”

Participants then answered a comprehension check question. If they answered incorrectly, they were re-directed to the instructions to review them again. Once they successfully completed the comprehension check, participants advanced to the main task. Figure 9A displays a sample trial for the speaker task. Participants viewed the 30 test videos as well as one attention check video. The order of the videos was randomized. Below the video on each trial, we provided the prompt “Which of the following sentences best describes the clip?” followed by the four description options with radio buttons. The order of the first three descriptions was randomized between participants, but the description with “made no difference” always came last. Participants had to view the video at least twice before making a selection. They were able to watch the video as many times as they liked, and chose to do so 2.2 times on average (SD = 0.6).

Analysis

Our model has four free parameters, θ which determines the amount of noise in the counterfactual simulations, σ which determines the softening parameter for the movement and uniqueness feature in the definition of “caused”, ν which determines the softening parameter in the definition of “made no difference”, and λ which controls the speaker optimality. We fit these parameters with a grid search, maximizing the likelihood of participant selections across trials under the corresponding distributions on utterances produced by the model. The range for θ was 0.5 to 1.6 in increments of 0.1. The range for σ and ν was 0 to 1 in increments of 0.1.⁸ The range for λ was 0.5 to 3.5 in increments of 0.25. We found an optimal value of 0.9 for θ , 0.4 for σ , 0.3 for ν , and 1.25 for λ .⁹

Alternative models. We compare our full model to two lesioned alternatives: a “No Pragmatics” model that removes the pragmatics component, and a “No Semantics and No Pragmatics” model that removes both components and computes a Bayesian ordinal regression instead, which directly maps from aspect values to utterance selections.

No Pragmatics This model removes the pragmatics component 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. This model is analogous to a “literal speaker”, who normalizes the semantic values across utterances instead of across scenarios (see Table 1b). We use the same parameter fitting procedure as in the full model with the same grid ranges. We found an optimal value of 1.0 for θ , 1.0 for σ , 1.0 for ν , and 2.75 for the temperature parameter of the softmax.¹⁰

No Semantics and No Pragmatics We fit a Bayesian ordinal regression with coefficients for each of the causal aspects, the movement feature, the uniqueness feature, random slopes and intercepts for each participant, and random intercepts for each trial. Because the predictors of the model are dependent on the noise parameter, θ , we fit one regression for each noise value sweeping across the same range of values we used for the other two models. The ordinal regression was written in Stan (Carpenter et al., 2017), and accessed with the `brms` package (Bürkner, 2017) in R (R Core Team, 2019). We assumed the following ordering of expressions (from weakest to strongest): “made no difference”, “affected”, “enabled”, “caused”. While this ordering is consistent with our semantics, the regression assumes a linear additive mapping from causal aspects to expressions, rather than the logical semantic mapping of the full model. The regression model has eight fixed effects, three thresholds determining the boundaries between each of the causal expressions, and five coefficients determining the weight of each of the aspects of causation along with the movement and

⁸The softeners, σ and ν , represent the probability of using the expression even if the condition they soften (movement, uniqueness, lack of how-cause) does not obtain. Because these are probabilities, their range is between 0 and 1.

⁹We also considered whether to use one or two levels of recursive reasoning for our pragmatic speaker (three levels and above became computationally prohibitive in the grid search). The data was slightly more likely under the best-performing level-2 model than under the level-1 model. So we report the results for the level-2 pragmatic speaker here.

¹⁰For the “No Pragmatics” model, both σ and ν are at the upper edge of the parameter range. We maintain the restriction on the range of these parameters given their interpretation as probabilities. If we widen the search ranges for these parameters enough to find a local optimum, the “No Pragmatics” model improves slightly, but the overall pattern of results is unchanged.

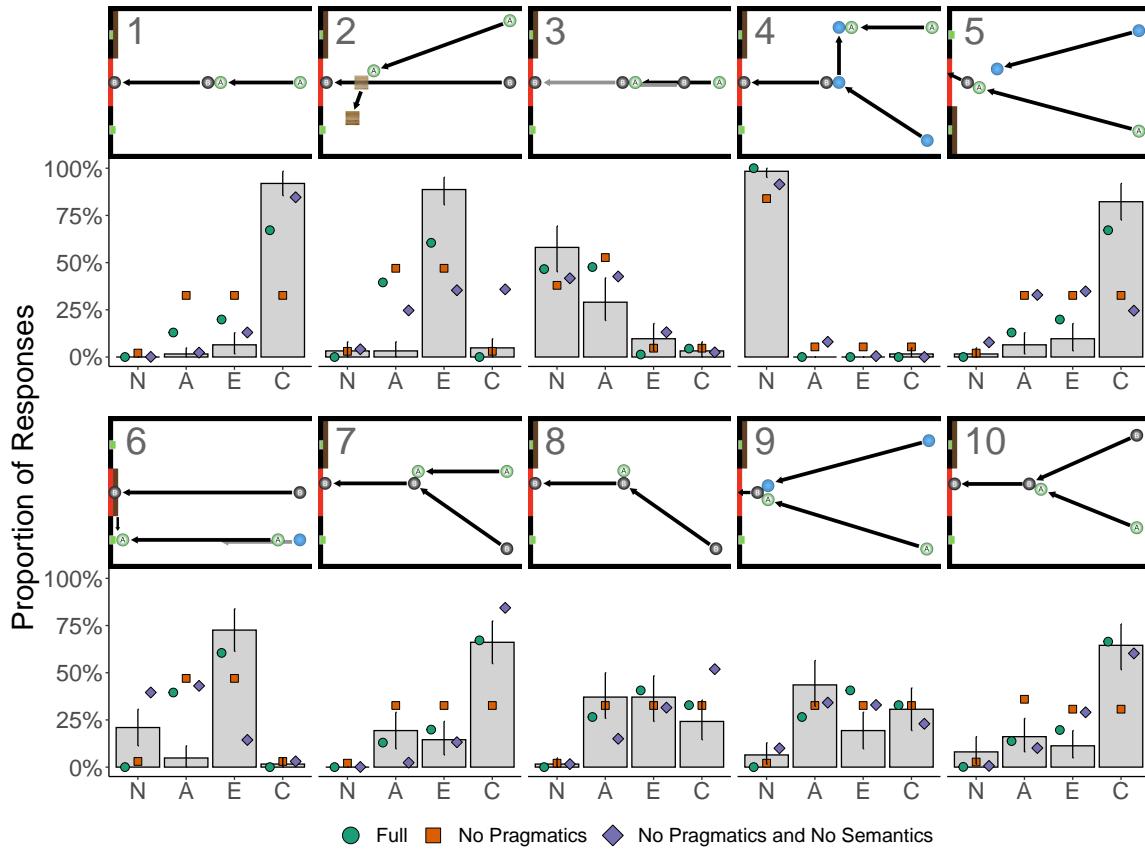


Figure 10. Experiment 2. Trial level model predictions for ten scenarios. In scenarios where participants strongly favor a particular utterance, the ‘Full Model’ captures this tendency, while the ‘No Pragmatics’ model assigns the same probability to all truthful utterances. Note: Error bars are bootstrapped 95% confidence intervals.

uniqueness feature. We found an optimal value of 1.0 for θ . Parameter estimates for this model are summarized in Table F1.

Results

Figure 10 shows participants’ selections for a subset of scenarios (Table E1 shows selections for all scenarios). In scenario 1, the classic Michottean case, participants strongly favor “caused”, while in scenario 2 where ball A clears the box from ball B’s path, participants strongly favor “enabled”. In scenario 3, the modal response is “made no difference”, though a substantial number of participants also selected “affected”. In scenario 4, nearly every participant selected “made no difference”. Scenarios 5 and 6 illustrate preemption scenarios. In scenario 5, where ball A directly contacts ball B, participants favor “caused”, while in scenario 6 where ball A hits a button opening the gate participants favor “enabled”. Interestingly, in scenario 6, a substantial number of participants selected “made no difference”. Scenarios 7 and 8 illustrate that movement matters, too. In scenario 7, where ball A is moving, participants strongly favor “caused”, but in scenario 8, which is identical except

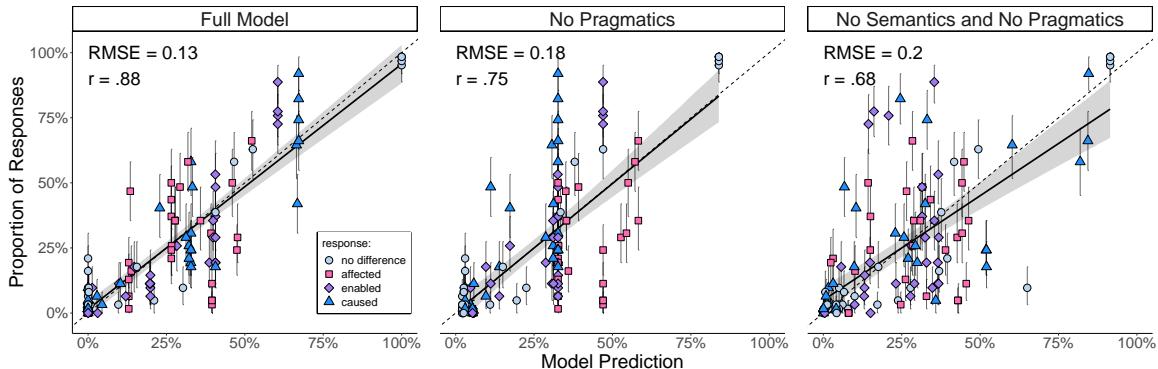


Figure 11. Experiment 2. Overall model performance for the “No Semantics and No Pragmatics” model, the “No Pragmatics” model, and the full model. Each point plots the model prediction for the probability of selecting a particular utterance on a given trial against the proportion of participants that selected that utterance on the same trial. The color and shape of the points indicates which utterance the point represents. There are thirty trials and four utterances per trial, so each panel contains 120 points. Overall, the full model performs best. The “No Pragmatics” model has a notable vertical line of responses around 33%. These reflect scenarios where multiple causal expressions are true, and so the “No Pragmatics” model ranks them all equally. The “No Semantics and No Pragmatics” model is more widely dispersed around the diagonal than the full model. *Note:* Error bars and regression bands show bootstrapped 95% confidence intervals.

ball A is stationary, participants are split between “affected”, “enabled”, and “caused”. In scenario 9, where ball A and the blue ball together collide with ball B driving it through the gate, participants are also split between “affected”, “enabled”, and “caused” (the modal response is “affected”). In scenario 10, where there is some uncertainty as to whether ball B would have gone through without ball A being present, participants still strongly favor “caused”.

Across these cases, the full model does the best job of capturing the data. In 7 out of 10 of the scenarios, the full model shows the closest match to the distribution of participant responses. When participants strongly favor a particular response (e.g. scenario 1, 2, and 4), the full model’s ability to select truthful and informative responses allows it to match this tendency. When participants show more variance in the expressions they select (e.g. scenario 3, 8, 9), the full model also matches this pattern. Both of the alternative models struggle with particular cases. The “No Pragmatics” model most clearly has issues with cases where participants favor “caused” (scenarios 1, 5, 8, 10). In these situations, “affected”, “enabled”, and “caused” are all true according to our semantics, so the “No Pragmatics” model rates them all equally. The “No Semantics and No Pragmatics” struggles with cases where participants favor “enabled” (scenario 2 and 6) as well as with scenario 5, likely because ball A is not a whether-cause in this case.

Figure 11 shows scatter plots of model predictions and aggregated participant responses for the full set of scenarios. 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 (“No Semantics and No Pragmatics”). In the “No Pragmatics”

model, we can see a large column around 33% on the x-axis, representing the cases where there are multiple true utterances and the model weighs them all equally. In the “No Semantics and No Pragmatics” model the responses are in general more broadly dispersed than the responses for the full model.

To further assess model fit, we performed 100 split-half cross validation runs for each model, splitting the data by trials.¹¹ Table 6 presents the results. The full model performs the best, followed by the “No Pragmatics” model, and the “No Semantics and No Pragmatics” model.

Discussion

In this experiment, we had participants select causal descriptions about what happened in various physical scenarios. This controlled setting allowed us to quantitatively compare model behavior to participant data. Overall, we see that each component of our model explains additional variance in participants’ responses. This suggests that causal reasoning, semantics, and pragmatics are all important for understanding how participants choose to describe what happened.

Though in general, the full model does a good job of capturing participant responses, there are some situations where its predictions are off. Notably, in cases like scenario 2 and scenario 6 where “enabled” is the dominant response, the full model over-weights the probability of “affected”, even though almost no participant used this utterance for these scenarios. One possibility is that how-cause is more important for the semantics of “affected” than the other aspects. Under our current definition, how-cause is a sufficient but not necessary condition for “affected” to be true. However, participants might feel that how-cause is both necessary and sufficient.

Another notable detail about scenario 6 is that a number of participants (around 20%) select “made no difference”. Scenario 6 is a situation of “preempted enablement”: ball A presses the button that moves the brown door out of the way allowing ball B to pass through the gate, but even if ball A hadn’t pressed the button, the blue ball would have done so. Interestingly, for the other situation where “enabled” is the dominant response

Table 6

Experiment 2. Speaker Experiment Split-Half Cross-Validation. The r column reports the median correlation coefficient on the test trials across the 100 cross-validation runs with 5% and 95% quantiles in brackets. The RMSE column reports the same for root mean square error. Δr reports the median difference in correlation coefficient between the Full model and the two alternative models, again with 5% and 95% quantiles in brackets. $\Delta RMSE$ reports the analogous difference in RMSE.

Model	r	Δr	RMSE	$\Delta RMSE$
Full Model	0.85 [0.77, 0.90]	–	0.14 [0.12, 0.18]	–
No Pragmatics	0.74 [0.63, 0.81]	0.11 [0.03, 0.18]	0.18 [0.15, 0.21]	0.03 [0.0, 0.07]
No Prag and No Sem	0.53 [0.27, 0.68]	0.32 [0.17, 0.54]	0.25 [0.20, 0.32]	0.10 [0.06, 0.16]

¹¹Because we split the data by trials, we excluded random intercepts for trials for the ordinal regression in cross-validation.

(scenario 2) and the other case of preemption (scenario 5), almost no participants respond “made no difference”.

Finally, it’s interesting that participants almost never selected the expression “enabled” in scenario 3. Most participants said that ball A either “made no difference” to the outcome, or that it “affected” ball B’s going through the gate – and this is predicted by the CSM. In contrast, the force dynamics model (Wolff, 2007) predicts that participants should say “enabled” here. In fact, the force configuration maps exactly onto the paradigmatic case for “enabled” as shown in Figure 1. One possible way of resolving this discrepancy is to assume that in many instances, there is some degree of uncertainty about whether the patient would have achieved the endstate by itself. For example, if there is friction then it’s possible that the patient would not have reached the endstate without the help of the agent. However, notice that this kind of reasoning is counterfactual and thereby outside of the purview of what the force dynamics model considers (although see Wolff et al., 2010, for an extension of the force dynamics model that incorporates some counterfactual machinery).

Experiment 3: Listeners inferring what happened

The previous experiment focused on the speaker side of communication. A speaker who saw what happened chooses what description to use. The listener is equally important. The RSA framework allows us to easily pivot our model to make predictions in the listener setting, too. In this third experiment, we examine what a listener can infer about the scenario given a causal description of what happened.

Methods

Stimuli. Figure 9B shows a sample trial for the listener task. Each trial consisted of a description (one of the four utterances from Experiment 1) and a pair of video scenarios. The scenarios were selected from the same set of videos used in Experiment 2. To select what pairs to show on a trial, we considered every possible video pair and evaluated their relative probability given each utterance under the pragmatic listener. For each utterance, we selected nine video pairs, varying the absolute difference in relative probabilities within each utterance. On some trials the model strongly preferred one video over the other, but on other trials, the model had a weak preference or no preference at all. In all trials, the utterance that the speaker chose was participants’ modal response in Experiment 2 for at least one of the scenarios shown in the trial.

Participants. We recruited 71 participants online via Mechanical Turk using Psi-turk (*age*: M = 37, SD = 10, *gender*: 19 female, 51 male, 1 non-binary, *race*: 40 White, 21 Black, 8 Asian, 2 unclear or no response). We removed 21 participants who failed to pass an attention check, leaving us with 50 participants for analysis.¹² The experiment took on average 31 minutes (SD = 13), and each participant was paid \$5.50.

Procedure. Participants first received instructions about the physical setting and the different objects in it. We then instructed participants on the speaker task from Experiment 2 and had them complete a comprehension check. After they passed the check,

¹²Attention check attrition was higher in this experiment. This could be in part because the check included two trials and was thus more stringent.

participants completed a short “training session” where they performed the speaker task for four trials. We then instructed participants on the listener task. Participants were told that rather than selecting a description for a video, they would now see a description that someone else chose. Their task was to indicate which of two scenarios they thought the hypothetical “describer” had seen based on the chosen description. Participants completed another comprehension check for this new task and then proceeded to the main phase of the experiment. Participants who failed either comprehension check were sent back to the instructions for the corresponding experiment section. Participants needed to pass all comprehension checks to proceed to the main phase of the experiment.

On each trial, participants saw which out of the four causal expression the describer had selected, indicated by a highlighted radio button next to the chosen expression (see Figure 9B). Participants were asked to answer the question: “Which video do you think the describer saw?” Below the prompt were two videos labeled “Video 1” on the left and “Video 2” on the right. Participants had to first watch the video on the left and then the video on the right. After they had watched each video once, a sliding scale appeared above the videos. The endpoints of the slider were labeled “Definitely Video 1” on the left, “Definitely Video 2” on the right, and the midpoint was labeled “Unsure”. The scale ranged from -50 to 50, though there were no numeric values visible to participants. To avoid anchoring effects, the slider handle was initially invisible and only appeared upon clicking on the slider. After a participant had viewed each video once and indicated a judgment on the slider, they could proceed to the next trial. Participants could replay videos as many times as they wanted (video play count $M = 1.13$, $SD = 0.41$).

Participants provided judgments on 36 trials and 2 attention checks. In one attention check, the describer’s chosen expression was “caused”, and one video illustrated a strong causal role of ball A while the other lacked any causal connection between ball A and ball B. In the second attention check, the chosen expression was “made no difference”, and again the comparison in the two scenarios was chosen to be maximally salient. Participants who put the slider on the opposite side of the correct video on at least one of the two attention checks were excluded from the analysis. The order of the trials and the position of the videos on each trial (left or right) was randomized between participants.

Analysis

We rescale participant responses to lie on the interval $[0, 1]$, where a zero response represents a judgment favoring the left scenario, and 1 represents a judgment favoring the right scenario. We compute the mean participant response for each trial.

With the model parameters that we fit in the speaker task, we use a level-1 pragmatic listener to compute a distribution on scenarios given each utterance. Each trial in our experiment consists of a given utterance and a pair of scenarios. To make a prediction for a particular trial, we take the probability of each scenario under the given utterance and apply a softmax with temperature parameter β . We fit β to minimize the squared error between model predictions and participant means.

Alternative Models. We include two alternative models analogous to the alternatives from Experiment 2. For the “No Pragmatics” model, we used a literal listener instead of a pragmatic listener to compute the distribution over scenarios given each utterance. For the “No Semantics and No Pragmatics” model, we take the predictions from the best-fitting

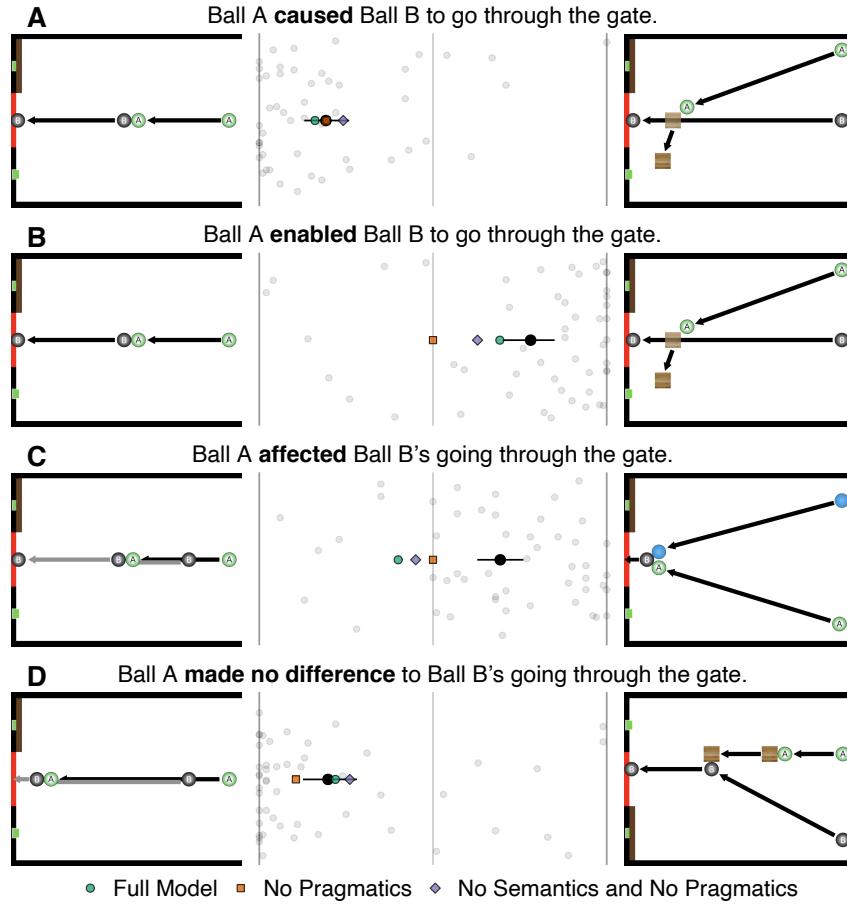


Figure 12. Experiment 3. Participant responses and model predictions for a selection of cases. Videos are illustrated on either end with the trial utterance listed above. Dark black dot represents participant mean selection with bootstrapped 95% confidence intervals. Light dots represent individual participant judgments. Colored shapes represent model predictions. **A)** All models do a good job capturing participant responses for the expression “caused”. **B)** For the same scenarios as in **A** when the expression is “enabled”, the “No Pragmatics” model fails to capture participant responses. Pragmatic inference allows the full model to infer the speaker intends to communicate “enabled” *but not “caused”* like participants do. **C)** All models fail to capture participants’ response tendency. Participants judge that “affected” better describes the scenario on the right, while the “Full Model” predicts that participants would prefer the scenario on the left. **D)** All models do a good job capturing participants’ preference for the left scenario given that “made no difference” was selected.

ordinal regression. We fit the predictions for both models using a softmax (with separate β parameters for each model).

Results

Figure 12 shows participant responses for four trials (Table G1 shows the responses for all trials). In Figure 12A, the speaker said that “Ball A **caused** ball B to go through the gate.” In one of the scenarios, ball A knocks into ball B and ball B goes through the gate (left side). In the other scenario, ball A knocks into a box pushing it out of ball B’s path and allowing ball B to go through the gate (right side). Participants judged that it was more likely that the speaker had referred to the scenario on the left, and all three models capture this preference. In Figure 12B, participants saw the same two scenarios, but this time the speaker said that “Ball A **enabled** ball B to go through the gate.” Here, participants strongly favor the scenario on the right. The full model matches participants’ responses more closely than the alternatives. Notably, the “No Pragmatics” model predicts that each scenario is equally likely because the utterance is true in both scenarios. The full model, like participants, draws the pragmatic inference that the speaker would have used the stronger utterance “caused” had they seen the scenario on the left, so they must have seen the scenario on the right.

In Figure 12C, the speaker said that “Ball A **affected** ball B’s going through the gate.” On the left, participants saw a case where ball B is headed toward the gate on its own, ball A comes up behind it and pushes it along. In this scenario, ball A is only a how-cause (but not a whether-cause or a sufficient-cause). On the right, they saw a case where ball A and the blue ball collide with ball B simultaneously driving it through the gate. In this scenario, ball A is a how-cause and a sufficient-cause. According to our semantics, “affected” is true for both scenarios, but “enabled” and “caused” are also true of the scenario on the right. Interestingly, participants strongly favored the scenario on the right. All of the models failed to capture this tendency. In this case, the full model has the greatest error because it again draws a pragmatic inference and thus leans toward the scenario on the left (because the speaker could have used a stronger expression had they wanted to refer to the video on the right). In Figure 12D, the speaker said that “Ball A **made no difference** to ball B’s going through the gate.” On the left, participants saw a scenario where ball A pushes ball B along after it is already headed to the gate. On the right, participants saw a scenario where ball A knocks into a box which then knocks into ball B and re-directs it through the gate. Participants strongly favored the scenario on the left here, and all models capture this pattern.

Figure 13 illustrates overall model performance for the full model and the two alternatives. The pattern of results is similar to the speaker experiment. The full model does the best job predicting participant responses, followed by the “No Pragmatics” model, and finally the “No Semantics and No Pragmatics Model”. For the “No Pragmatics” model, there is again a column of responses near the middle. This is the case because the “No Pragmatics” can’t distinguish between scenarios on trials in which the given utterance is true of both scenarios.

As with the speaker experiment, we ran 100 split-half cross-validations to compare model fits. The results are summarized in Table 7. Again we see that the full model outperforms the alternatives. This time, the difference in performance between the full model and the “No Pragmatics” model is smaller than in Experiment 2.

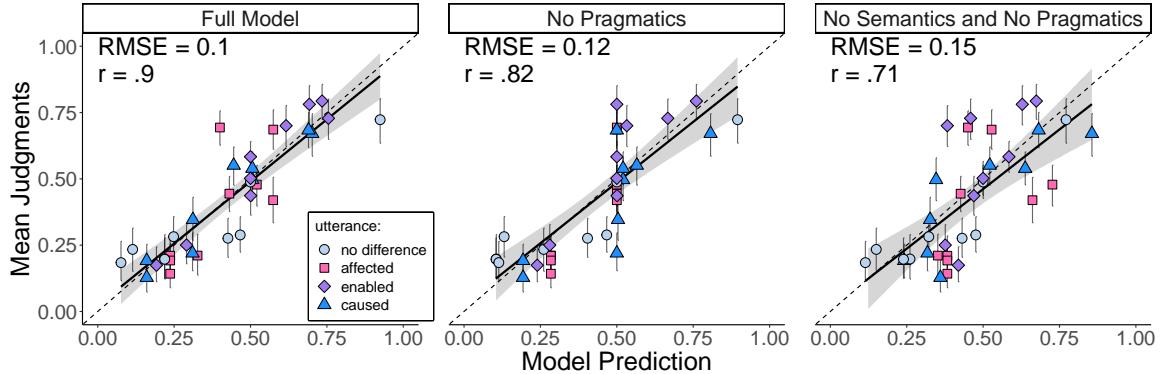


Figure 13. Experiment 3. Overall performance for the full model and alternatives in the listener experiment. Each point represents the model prediction for a particular trial compared against the mean participant response on that trial. The utterance type is indicated by the color and shape of the point. Note: Error bars and regression bands show bootstrapped 95% confidence intervals.

Table 7

Experiment 3. Listener Experiment Split-Half Cross-Validation. The r column reports the median correlation coefficient on the test trials across the 100 cross-validation runs with 5% and 95% quantiles in brackets. The RMSE column reports the same for root mean square error. Δr reports the median difference in correlation coefficient between the full model and the two alternative models, again with 5% and 95% quantiles in brackets. ΔRMSE reports the analogous difference in RMSE.

Model	r	Δr	RMSE	ΔRMSE
Full Model	0.91 [0.84, 0.95]	–	0.10 [0.07, 0.12]	–
No Pragmatics	0.83 [0.76, 0.90]	0.07 [0.0, 0.14]	0.13 [0.10, 0.15]	0.03 [0.0, 0.05]
No Prag and No Sem	0.72 [0.60, 0.81]	0.18 [0.09, 0.31]	0.16 [0.14, 0.18]	0.06 [0.03, 0.09]

Discussion

In this experiment we demonstrate that our model explains patterns of participant behavior in the listener setting. In terms of model performance, the pattern of results is very similar to what we saw in the speaker experiment (Experiment 2). The full model performs better than both of the alternatives in the cross-validation. This time, the performance difference between the full model and the “No Pragmatics” was smaller. However, the full model holds a qualitative advantage over the “No Pragmatics” model in that it correctly predicts participant responses on trials that involve scalar implicatures. In trials like the one presented in Figure 12B, the full model infers that because the speaker used a weaker utterance, they intended to communicate that the ball “enabled” but didn’t “cause” the outcome. This behavior is consistent with the semantic and pragmatic assumptions of our model which we validated in Experiment 1.

Certain cases still pose a challenge for our model. This is most clearly apparent in the trial depicted in Figure 12C. Though all the models fail to capture the general pattern of participant responses in this case, the full model is especially off. Because the full model

believes stronger utterances are true of the scenario on the right, it favors the scenario on the left for which “affected” (and to some extent “made no difference”) are the only true utterances. Interestingly, in the speaker experiment, a plurality of participants favored “affected” for the scenario on the right (see scenario 9 in Figure 10).

For the listener task, we can also fit an empirical model based on participant responses in the speaker task. For each trial, this model takes the proportion of participants that responded with the given utterance on the two scenarios in the speaker experiment and then normalizes these two values using a softmax function (with a fitted β parameter). For example, in Figure 12A, the model would take the proportion of participants from Experiment 2 who responded “caused” in the scenario on the left and on the right, and then run this pair of values through a fitted softmax. This empirical model captures participants’ inferences in the listener experiment very well with $r = .95$, RMSE = 0.07. The fact that the empirical model performs so well suggests that the listener task doesn’t introduce many additional factors that go beyond what participants need to do in the speaker task. As our model suggests, a listener can infer what happened simply by considering the extent to which they would have used a given expression in the different situations (cf. Kirfel, Icard, & Gerstenberg, 2022).

General Discussion

Causality permeates our everyday language in ways both subtle and pronounced. In this paper, we developed a computational model of people’s use of different causal expressions including “caused”, “enabled”, “affected”, and “made no difference” that draws together insights from philosophy, linguistics, and psychology. Our model consists of three modules: a causal knowledge module, a semantics module, and a pragmatics module. The *causal knowledge module* computes a representation of different causal aspects of a scene by simulating the consequences of different counterfactual interventions. The *semantics module* defines different causal expressions in terms of logical combinations of these causal aspects. The *pragmatics module* then computes which utterance to use (or scenario to infer) based on principles of rational communication. Together these three components offer an account of how people choose and interpret causal expressions.

We tested our model in a series of experiments. In an initial set of psycholinguistic experiments, we validated the model’s semantics and provided evidence that people draw pragmatic inferences as predicted by our model. Across two experiments, we asked participants to rate the acceptability of a series of sentences that tested our model semantics (Experiment 1B) and pragmatics (Experiment 1C). We found that the qualitative pattern of participants’ judgments was consistent with our model’s predictions.

We followed this qualitative validation of our model with two quantitative tests. In Experiment 2, participants took the role of a speaker. They viewed video scenarios of physical interactions and chose what sentence best described what happened. In Experiment 3, participants took the role of a listener. Their task was to infer which of two videos a speaker had seen based on the causal expression they had selected. Our model captured participant judgments well in both these tasks. To assess model performance, we compared the full model with two lesioned alternative models, removing either only the pragmatics module, or both the pragmatics and the semantics modules. In both experiments, we found that

the full model outperformed the alternatives in cross-validation, suggesting that all three components of the model are important for understanding how people perform these tasks.

Our work here establishes a framework for modeling the relationship between people’s causal representations of the world and their causal language. The implementation of this model requires various assumptions. In the remainder, we will consider some limitations of our implementation that suggest potential avenues for future research, discussing the three different modules one by one: causal knowledge, semantics, and pragmatics.

Causal Knowledge

The causal knowledge module is based on the CSM (Gerstenberg et al., 2021). The CSM was developed to model causal judgments in intuitive physical settings, and that was the domain we focused on for quantitative modeling here. But people use causal language across a wide variety of domains, and a more comprehensive model would capture these different kinds of causal interactions. New questions of implementation arise when generalizing the tools of the CSM to new settings. As a starting point, the CSM requires a generative model which supports counterfactual simulation in the domain of interest. Recent work has begun to implement these models for social causal interactions (Sosa, Ullman, Tenenbaum, Gershman, & Gerstenberg, 2021; Wu, Sridhar, & Gerstenberg, 2022; Wu et al., 2023). Rather than relying on noisy physics simulators, these approaches use models of planning and rational action (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016). They describe how agents select actions by weighing costs and benefits and how they update their beliefs about the world. These models also support inferences about the latent beliefs and desires that give rise to an observed set of actions (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017).

With a new set of generative models, additional questions arise about how to implement counterfactual tests in the CSM. When testing for whether-causation in the physical domain, removing the candidate cause is a straightforward way to assess counterfactual necessity. In the social domain, the appropriate counterfactual operation is less obvious. One possible contrast is removing the agent altogether, but this might not be the standard counterfactual that people rely on when assessing social causal roles. For instance in the law, the reasonable person test (Jackson, 2013) is a counterfactual test that prompts jurors and judges to compare the behavior of a defendant to an imagined reasonable person (cf. Gerstenberg et al., 2018). The agent isn’t removed from the setting; rather their mental states are manipulated such that they meet a standard of reasonableness. Generative models of social cognition can support these kinds of counterfactuals, though defining precisely what kind of counterfactual operations are required to make someone reasonable is not obvious *a priori* and requires further research at the intersection of social cognition and law (Tobia, 2018).

Though the physical and social world are two of the most pervasive domains of causal thinking, people’s causal thought ranges further still, supporting judgments about topics as diverse as human biology (“Eating the ice cream gave me stomach ache.”), weather (“The heavy rains caused the flowers to bloom.”), and political-economic events (“The war in Ukraine made gas prices go up.”). In principle, the CSM could generalize to model causal thinking in such domains by operating over increasingly abstract generative models (Beckers & Halpern, 2019; Ho et al., 2022; Shin & Gerstenberg, 2023). Critically, these

models need not maintain a detailed representation of the processes that support causal reasoning in the physical and social domains. Laypeople making judgments about the causal relationships in complex phenomena likely operate with relatively simplified intuitive models about the mechanisms that underlie those relationships (Rozenblit & Keil, 2002). Even experts in these fields need to make various simplifying assumptions in order to model complex phenomena at the level of, for instance, economic systems. As long as the model supports counterfactual simulation, it also supports the ability to reason about cause and effect.

Semantics

In our model, we propose a semantics for three periphrastic causatives: “caused”, “enabled”, and “affected”. According to our semantics, these expressions exist on a hierarchy of informativeness, where “caused” is the most informative, “enabled” is in the middle, and “affected” is the least informative. In Experiment 1, we showed that participants broadly affirm this informative hierarchy for causal statements describing a wide range of causal phenomena. In Experiments 2 and 3 we further demonstrated that, with a specific semantics defined in terms of the aspects of causation that implies this hierarchy, our model is able to make linguistic choices that match participant judgments in a causal physical domain.

Though our semantics does a good job of capturing participant behavior in these experimental contexts, there are limits to its generality. In different domains it is not clear necessarily how to define the different aspects of causation. Relatedly, the particular inferences that one might draw given one of our causal expressions might vary based on the domain. In our physical setting, if a person hears that “Ball A enabled ball B to go through the gate, but it didn’t cause it”, it may be natural to infer that ball A was a whether-cause (or sufficient-cause) but not a how-cause. On the other hand if one hears that “The Sackler’s greed enabled the opioid epidemic, but it didn’t cause it” one might not draw the same inference. It is not immediately clear what it would mean for the Sackler’s greed to be a how-cause of the opioid epidemic. It may be apparent that the speaker is making some claim about the directness (or importance) of the Sackler’s causal role, but what exactly that claim amounts to is not specified by our model in its current form. To investigate how people infer specific details of a speaker’s causal representation based on their choice of causal expression, we will need more developed causal models that characterize how the different aspects are implemented in the different domains.

Though the specific semantic definitions may be somewhat flexible based on the domain, the more general claim that these three causal expressions exist on an informative hierarchy appears more robust. Experiment 1 illustrates that across many different causal contexts, participants are willing to endorse the hierarchy of informativeness implied by our semantics. This suggests that this hierarchical semantic relationship among the three verbs holds more generally, at least in contexts where all three verbs are acceptably substitutable for one another. However, there are limitations to the generality of this observation, as one can construct counterexamples to this semantic hierarchy. Wolff (2007, p. 84) describes the following example:

1. A cold wind caused him to close the window.
2. ?A cold wind enabled him to close the window.

According to our semantics, whenever one can say “caused”, one should also be able to say “enabled” as well. But this doesn’t seem right here. While “caused” seems perfectly fine, “enabled” doesn’t sound good. This example suggests that there are further constraints on the use of “enabled” that our current semantics doesn’t account for. One possible explanation is that “enabled” imposes additional restrictions on the causal structure being described. According to Sloman et al. (2009) “enabled” implies a causal structure where the subject is necessary for the outcome, but another causal factor exists as well. Here, no such additional factor is clear from the context, and thus the sentence appears like a misuse of “enabled” (see also Cao, Geiger, Kreiss, Icard, & Gerstenberg, 2023).

In addition to better understanding the nuanced meanings of the causal expressions we currently model, future work needs to expand the set of expressions. Beyond “caused”, “enabled”, and “affected”, linguists and psychologists have analyzed a wide variety of periphrastic causatives. Their analyses suggest fruitful avenues for expanding the causal lexicon in our framework. Nadathur and Lauer (2020) explore the semantics of “make”, highlighting the importance of sufficiency in its meaning. Sufficiency is already implemented in the CSM, and it would be straightforward to provide a definition in the causal physical setting and then test that definition alongside our other expressions. Another interesting verb which has been studied by linguists is “let”. Talmy (1988) analyzes “let” with a force dynamics configuration where a stronger antagonist is withdrawn thereby allowing the agonist to proceed with its prior tendency. This configuration bears similarity to Figure 4 scenario 2, where ball A knocks a box out of the way allowing ball B to pass through the gate. In our experiment, participants strongly favored “enabled” for this case (see Figure 10), and “let” does seem semantically similar to “enabled”. Wolff and Song (2003) identify a collection of “enable-type” verbs with closely related meanings which include “enable” and “let” as well as “allow” and “help”. Our framework provides the potential to make explicit hypotheses about the differences (and similarities) in meanings among these verbs and then test those hypotheses quantitatively. Pushing even further in this direction, we aim to ultimately define lexical causative verbs such as “break”, “melt”, and “kill” in this framework.

An important limitation of our work is that, so far, the model only applies to a set of English causatives. In the future, we would like to expand this approach to other languages as well. Cross-linguistic research has revealed interesting similarities but also important differences in the ways that causation is expressed and understood across languages (Beller, Song, & Bender, 2009; Bender & Beller, 2011; Comrie, 1976; Haspelmath, 2016; Wolff, Jeon, & Li, 2009; Wolff, Klettke, Ventura, & Song, 2005). The psychological representations that underlie the use of a word in one language and its translated equivalent in another might not be the same. For example, Klettke and Wolff (2003) find that English and German experimental participants differ in their tendencies to describe the same scenarios using “cause” (German “verursachen”) or “enable” (“ermöglichen”). Our model provides a framework to explore the psychological underpinnings of these differences. Ultimately, this cross-linguistic examination would provide a fuller picture of how people use causal language around the world.

Pragmatics

In our model, we used the Rational Speech Acts (RSA) framework to model participants’ use of scalar implicature in our tasks. In the literature on scalar implicature, linguists

often distinguish between lexicalized scales and ad hoc scales (Degen, 2015; Hirschberg, 1985). Whereas in a lexicalized scale, the use of a weaker utterance always invites the stronger utterance as an alternative (as is often suggested for “some” and “all”), in an ad hoc scale, additional context may be required to make the comparison salient. It is difficult to assess whether the scale we’ve identified in this work is lexicalized or ad hoc. The setup in our experiments makes the relevant contrast very clear. While this does not rule out the possibility that people automatically consider “cause” as an alternative when hearing “enable” in natural speech, we would need different methods to test what alternatives naturally come to people’s minds.

The existence of ad hoc scales however highlights the role of context and comparison in pragmatic speech, and raises questions about how the contrast set might impact our analysis here. To some extent, the effects we’ve observed are shaped by the set of alternative utterances that we had participants consider. Naturally, we might wonder what would happen if we changed the set of alternatives. For example, would people use “caused” in our scenarios differently when “made” was an alternative? Hand-in-hand with our future studies of the semantics of other causative constructions come questions about the pragmatics. If we define additional causative constructions, how does pragmatic inference drive the choice of one or another given a particular context? What if we contrast lexical causatives with periphrastic ones? The linguistics literature has explored extensively the conditions under which it is appropriate to use a lexical causative or a periphrastic alternative (Fodor, 1970; Katz, 1970; McCawley, 1978; Shibatani, 1976), though sometimes these examinations are performed by individual linguists on a small number of examples. Our model provides a new framework for examining these questions, allowing us to formulate explicit hypotheses about the semantics and pragmatics of causal verbs.

These directions for future study highlight the inter-relatedness of the semantic and pragmatic modules of our model. Though in our work here, these modules are neatly separated, studies in linguistics reveal that the assumed separation between these capacities is often not so clear-cut (Börjesson, 2014). Addressing this observation, Bergen, Levy, and Goodman (2016) develop an RSA framework where pragmatic inference extends to the semantic content of the expressions themselves. Rather than assuming the semantic content of utterances is fixed across contexts, this approach assumes that interlocutors maintain a level of lexical uncertainty over the semantics of different utterances. In addition to inferring which utterance a speaker would choose in context, a listener also infers what exactly the speaker means by an utterance (see also Potts, Lassiter, Levy, & Frank, 2015). Extending our model in this way could expand its explanatory power and help address problem cases like the one from Wolff (2007) which seems to suggest additional semantic properties of “enabled” that we have not accounted for.

Conclusion

Causation is complex and multi-faceted. There are many ways in which one event can make a difference to another, and our language provides us with a limited set of expressions to communicate what happened. Did Charlie cause Phil’s success, or merely enable it? To capture how people use and understand various causal expressions, we developed and tested the *counterfactual simulation model of causal language*. The model assumes that people form a causal representation of what happened by paying attention to the way in

which the candidate cause brought about the outcome. A cause can make a difference to whether and how an outcome happened, and this is revealed through simulations of what would have happened in relevant counterfactual situations. The meaning of different causal expressions, such as “caused”, “enabled”, “affected”, and “made no difference” is then defined in terms of these aspects of causation, and by incorporating pragmatic principles of rational communication, the model accounts for how speakers choose what words to use, and how listeners infer what happened. By drawing on insights from philosophy, linguistics, and psychology, the work presented here brings us one step closer to understanding how people communicate about causality.

Acknowledgments

Many thanks to Erin Bennett for her many contributions in the early stages of this project! We also thank Bella Fascendini, Beth Levin, Dan Lassiter, David Rose, Judith Degen, Mike Frank, Thomas Icard, as well as the members of the Causality in Cognition Lab (CiCL) for helpful feedback and discussion. TG was supported by a research grant from the Stanford Institute for Human-Centered Artificial Intelligence (HAI).

References

- Aronson, J. L. (1971). On the grammar of ‘cause’. *Synthese*, 22(3), 414–430.
- Baglini, R., & Siegal, E. A. B.-A. (2021). Modelling linguistic causation. *manuscript, Aarhus University and Hebrew University of Jerusalem*.
- Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4), 0064.
- Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45), 18327–18332.
- Beckers, S. (2021). Causal sufficiency and actual causation. *Journal of Philosophical Logic*, 50(6), 1341–1374.
- Beckers, S., & Halpern, J. Y. (2019). Abstracting causal models. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 2678–2685).
- Beller, S., Song, J., & Bender, A. (2009). Weighing up physical causes: Effects of culture, linguistic cues and content. *Journal of Cognition and Culture*, 9(3), 347–365.
- Bender, A., & Beller, S. (2011). Causal asymmetry across cultures: assigning causal roles in symmetric physical settings. *Frontiers in psychology*, 2, 231.
- Bergen, L., Levy, R., & Goodman, N. (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, 9, 1–83.
- Börjesson, K. (2014). The semantics-pragmatics controversy. In *The semantics-pragmatics controversy*. de Gruyter.
- Bunzl, M. (1980). Causal preemption and counterfactuals. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 37(2), 115–124.
- Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 1–28.
- Cao, A., Geiger, A., Kreiss, E., Icard, T., & Gerstenberg, T. (2023). A semantics for causing, enabling, and preventing verbs using structural causal models.
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1).
- Chang, W. (2009). Connecting counterfactual and physical causation. In *Proceedings of the 31th annual conference of the cognitive science society* (pp. 1983–1987). Cognitive Science Society, Austin, TX.
- Chater, N., & Oaksford, M. (2013). Programs as causal models: Speculations on mental programs and mental representation. *Cognitive Science*, 37(6), 1171–1191.
- Cheng, P. W., & Novick, L. R. (1991). Causes versus enabling conditions. *Cognition*, 40, 83–120.
- Comrie, B. (1976). The syntax of causative constructions: cross-language similarities and divergences. In *The grammar of causative constructions* (pp. 259–312). Brill.
- Cruse, D. A. (1972). A note on english causatives. *Linguistic inquiry*, 3(4), 520–528.
- Degen, J. (2015). Investigating the distribution of some (but not all) implicatures using corpora and web-based methods. *Semantics and Pragmatics*, 8(11), 1–55.
- Degen, J. (2023). The rational speech act framework. *Annual Review of Linguistics*, 9,

- 519–540.
- De Leeuw, J. R. (2015). jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods*, 47, 1–12.
- Dowe, P. (2000). *Physical causation*. Cambridge, England: Cambridge University Press.
- Fair, D. (1979). Causation and the flow of energy. *Erkenntnis*, 14(3), 219–250.
- Flanigan, J. (2014). A defense of compulsory vaccination. In *Hec forum* (Vol. 26, pp. 5–25).
- Fodor, J. A. (1970). Three reasons for not deriving “kill” from “cause to die”. *Linguistic Inquiry*, 1(4), 429–438.
- Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford review*, 5, 5–15.
- Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084), 998–998.
- Garvey, C., & Caramazza, A. (1974). Implicit causality in verbs. *Linguistic inquiry*, 5(3), 459–464.
- Gerstenberg, T. (2022). What would have happened? counterfactuals, hypotheticals and causal judgements. *Philosophical Transactions of the Royal Society B*, 377(1866), 20210339.
- Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, 128(5), 936–975.
- Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017). Eye-tracking causality. *Psychological Science*, 28(12), 1731–1744.
- Gerstenberg, T., & Stephan, S. (2021). A counterfactual simulation model of causation by omission. *Cognition*, 216, 104842.
- Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. In M. Waldmann (Ed.), *Oxford handbook of causal reasoning* (pp. 515–548). Oxford University Press.
- Gerstenberg, T., Ullman, T. D., Nagel, J., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2018). Lucky or clever? From expectations to responsibility judgments. *Cognition*, 177, 122–141.
- 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). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11), 818–829.
- Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, 5(1), 173–184.
- Grice, H. P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.
- 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.
- Hall, N. (2004). Two concepts of causation. In J. Collins, N. Hall, & L. A. Paul (Eds.), *Causation and counterfactuals*. MIT Press.
- Hall, N. (2007). Structural equations and causation. *Philosophical Studies*, 132, 109–136.
- 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.
- Hartshorne, J. K. (2013). What is implicit causality? *Language, Cognition and Neuroscience*, 29(7), 804–824.
- Haspelmath, M. (2016). Universals of causative and anticausative verb formation and the spontaneity scale. *Lingua Posnaniensis*, 58(2), 33–63.
- Henne, P., & O'Neill, K. (2022). Double prevention, causal judgments, and counterfactuals. *Cognitive Science*, 46(5), e13127.
- Hirschberg, J. B. (1985). *A theory of scalar implicature (natural languages, pragmatics, inference)*. (Doctoral dissertation, University of Pennsylvania)
- Hitchcock, C. (1995). Salmon on explanatory relevance. *Philosophy of Science*, 62(2), 304–320.
- Hitchcock, C. (2009). Structural equations and causation: six counterexamples. *Philosophical Studies*, 144(3), 391–401.
- Hitchcock, C., & Knobe, J. (2009). Cause and norm. *Journal of Philosophy*, 11, 587–612.
- Ho, M. K., Abel, D., Correa, C. G., Littman, M. L., Cohen, J. D., & Griffiths, T. L. (2022). People construct simplified mental representations to plan. *Nature*, 606(7912), 129–136.
- Hobbs, J. R. (2005). Toward a useful concept of causality for lexical semantics. *Journal of Semantics*, 22(2), 181–209.
- Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, 161, 80–93.
- Jackson, C. (2013). Reasonable persons, reasonable circumstances. *San Diego L. Rev.*, 50, 651–706.
- 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(10), 589–604.
- Jensen, D. (2019). Overcoming the poverty of mechanism in causal models. In S. Kleinberg (Ed.), *Time and causality across the sciences*. Cambridge University Press.
- Johnson-Laird, P. N. (1989). *Mental models*. The MIT Press.
- Katz, J. J. (1970). Interpretative semantics vs. generative semantics. *Foundations of language*, 220–259.
- Kaufmann, S. (2013). Causal premise semantics. *Cognitive Science*, 37(6), 1136–1170.
- Kirfel, L., Icard, T., & Gerstenberg, T. (2022). Inference from explanation. *Journal of Experimental Psychology: General*, 151(7), 1481–1501.
- Klettke, B., & Wolff, P. (2003). Differences in how english and german speakers talk and reason about cause. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 25).
- Kubricht, J. R., Holyoak, K. J., & Lu, H. (2017). Intuitive physics: Current research and controversies. *Trends in Cognitive Sciences*, 21(10), 749–759.
- Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science*, 47, 1036–1073.
- Langenhoff, A. F., Wiegmann, A., Halpern, J. Y., Tenenbaum, J. B., & Gerstenberg, T. (2021). Predicting responsibility judgments from dispositional inferences and causal attributions. *Cognitive Psychology*, 129, 101412.
- Levin, B., & Hovav, M. R. (1994). A preliminary analysis of causative verbs in english.

- Lingua*, 92, 35–77.
- Lewis, D. (1973). Causation. *The Journal of Philosophy*, 70(17), 556–567.
- Livengood, J. (2011). Actual causation and simple voting scenarios. *Nous*, 1–33.
- Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, 61(4), 303–332.
- Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, 67(1), 1–25.
- Mackie, J. L. (1974). *The cement of the universe*. Oxford: Clarendon Press.
- Malm, H. M. (1989). Killing, letting die, and simple conflicts. *Philosophy & public affairs*, 238–258.
- Matsumoto, Y. (1997). Scales, implicatures, and in fact, if not, and let alone constructions. *Studies in English*, 685–699.
- Mayol, L., & Castroviejo, E. (2013). How to cancel an implicature. *Journal of Pragmatics*, 50(1), 84–104.
- Mayrhofer, R., & Waldmann, M. R. (2016). Causal agency and the perception of force. *Psychonomic Bulletin & Review*, 23(3), 789–796.
- McCawley, J. D. (1978). Conversational implicature and the lexicon. *Syntax and semantics*, 9, 245–259.
- 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), 1–37.
- Niemi, L., Hartshorne, J., Gerstenberg, T., Stanley, M., & Young, L. (2020). Moral values reveal the causality implicit in verb meaning. *Cognitive Science*, 44(6), e12838.
- Paul, L. A., & Hall, N. (2013). *Causation: A user's guide*. Oxford University Press.
- Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, England: Cambridge University Press.
- Potts, C., Lassiter, D., Levy, R., & Frank, M. C. (2015). Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. *Journal of Semantics*, 33(4), 755–802.
- R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.
- Rodriguez-Arias, D., Rodriguez Lopez, B., Monasterio-Astobiza, A., & Hannikainen, I. R. (2020). How do people use ‘killing’, ‘letting die’ and related bioethical concepts? contrasting descriptive and normative hypotheses. *Bioethics*, 34(5), 509–518.
- Rose, D., Sievers, E., & Nichols, S. (2021). Cause and burn. *Cognition*, 207(104517).
- Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive science*, 26(5), 521–562.
- Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press, Princeton NJ.
- Schaffer, J. (2000). Causation by disconnection. *Philosophy of Science*, 67(2), 285.
- Schaffer, J. (2013). Causal contextualisms. In *Contrastivism in philosophy* (pp. 43–71).

- Routledge.
- Shibatani, M. (1976). The grammar of causative constructions: A conspectus. In *The grammar of causative constructions* (pp. 1–40). Brill.
- Shin, S. M., & Gerstenberg, T. (2023). Learning what matters: Causal abstraction in human inference. In M. B. Goldwater, F. Anggoro, B. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society*.
- Sloman, S. A. (2005). *Causal models: How people think about the world and its alternatives*. Oxford University Press, USA.
- 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.
- Smith, C. S. (1970). Jespersen's 'move and change' class and causative verbs in english. *Linguistic and literary studies in honor of Archibald A. Hill*, 2, 101–109.
- Smith, K. A., Hamrick, J. B., Sanborn, A. N., Battaglia, P. W., Gerstenberg, T., Ullman, T. D., & Tenenbaum, J. B. (in press). Probabilistic models of physical reasoning. In T. L. Griffiths, N. Chater, & J. B. Tenenbaum (Eds.), *Reverse engineering the mind: Probabilistic models of cognition*.
- Smith, K. A., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in Cognitive Science*, 5(1), 185–199.
- Sosa, F. A., Ullman, T., Tenenbaum, J. B., Gershman, S. J., & Gerstenberg, T. (2021). Moral dynamics: Grounding moral judgment in intuitive physics and intuitive psychology. *Cognition*, 217, 104890.
- Spirites, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search*. The MIT Press.
- Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, 12(1), 49–100.
- Thomson, J. J. (1976a). A defense of abortion. In *Biomedical ethics and the law* (pp. 39–54). Springer.
- Thomson, J. J. (1976b). Killing, letting die, and the trolley problem. *The Monist*, 59(2), 204–217.
- Tobia, K. P. (2018). How people judge what is reasonable. *Alabama Law Review*, 70, 293–359.
- Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, 21(9), 649–665.
- White, P. A. (2014). Singular clues to causality and their use in human causal judgment. *Cognitive Science*, 38(1), 38–75.
- Wierzbicka, A. (1975). Why "kill" does not mean "cause to die": the semantics of action sentences. *Foundations of language*, 13(4), 491–528.
- 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.
- 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.
- Wolff, P., Jeon, G.-H., & Li, Y. (2009). Causers in english, korean, and chinese and the

- individuation of events. *Language and Cognition*, 1(2), 167–196.
- Wolff, P., Klettke, B., Ventura, T., & Song, G. (2005). Expressing causation in english and other languages. In W. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (Eds.), *Categorization inside and outside the laboratory: Essays in honor of Douglas L. Medin* (p. 29-48). American Psychological Association.
- Wolff, P., & Song, G. (2003). Models of causation and the semantics of causal verbs. *Cognitive Psychology*, 47(3), 276–332.
- Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford, England: Oxford University Press.
- Woodward, J. (2006). Sensitive and insensitive causation. *The Philosophical Review*, 115(1), 1–50.
- Woodward, J. (2021). *Causation with a human face: Normative theory and descriptive psychology*. Oxford University Press.
- Wu, S., Sridhar, S., & Gerstenberg, T. (2022). That was close! a counterfactual simulation model of causal judgments about decisions. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).
- Wu, S. A., Sridhar, S., & Gerstenberg, T. (2023). A computational model of responsibility judgments from counterfactual simulations and intention inferences. In M. B. Goldwater, F. Anggoro, B. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society*.
- Zhou, L., Smith, K. A., Tenenbaum, J. B., & Gerstenberg, T. (2023). Mental Jenga: A counterfactual simulation model of causal judgments about physical support. *Journal of Experimental Psychology: General*.

Appendix A
Sentence frames from norming study (Experiment 1A).

Table A1

All sentence frames from norming study. Included sentences had median ratings for all verbs above the midpoint of the scale. Excluded sentences had median ratings at the midpoint or below for at least one causal expression.

Included Sentences
1. The dry weather ____ the wild fire.
2. The CEO's decision ____ the outcome.
3. The new technology ____ the change.
4. The Sackler's greed ____ the opioid epidemic.
5. The sunny weather ____ the tree's growth.
6. More stipends ____ the increase in student admissions.
7. The sun ____ the drying of the clothes.
8. Metastasis ____ cell growth.
9. Diversification ____ new monetary policies.
10. The algae buildup in the ocean ____ the migration of certain species of fish.
Excluded Sentences
11. The collapse of Lehman Brothers ____ the financial crisis.
12. The zoning restrictions ____ the housing shortage.
13. The breaking of the dam ____ the flood.
14. Janelle's working hard ____ her success.
15. The new traffic signs ____ the decrease in fatalities.
16. The striker deflecting the ball ____ the goal.
17. Turning off life support ____ the patient's death.
18. Deforestation ____ wildlife displacement.
19. Erosion ____ density loss.
20. The construction at the intersection ____ the traffic in the vicinity.
Attention Checks
1. The lightening strike caused the fire.
2. The crank caused him to open the window.
3. Receiving the loan enabled her to buy the house.
4. The cold breeze enabled him to close the window.
5. The distracting noise affected his performance.
6. The earthquake affected the building to fall.

Appendix B
Experiment 1B Affected → Enabled Items

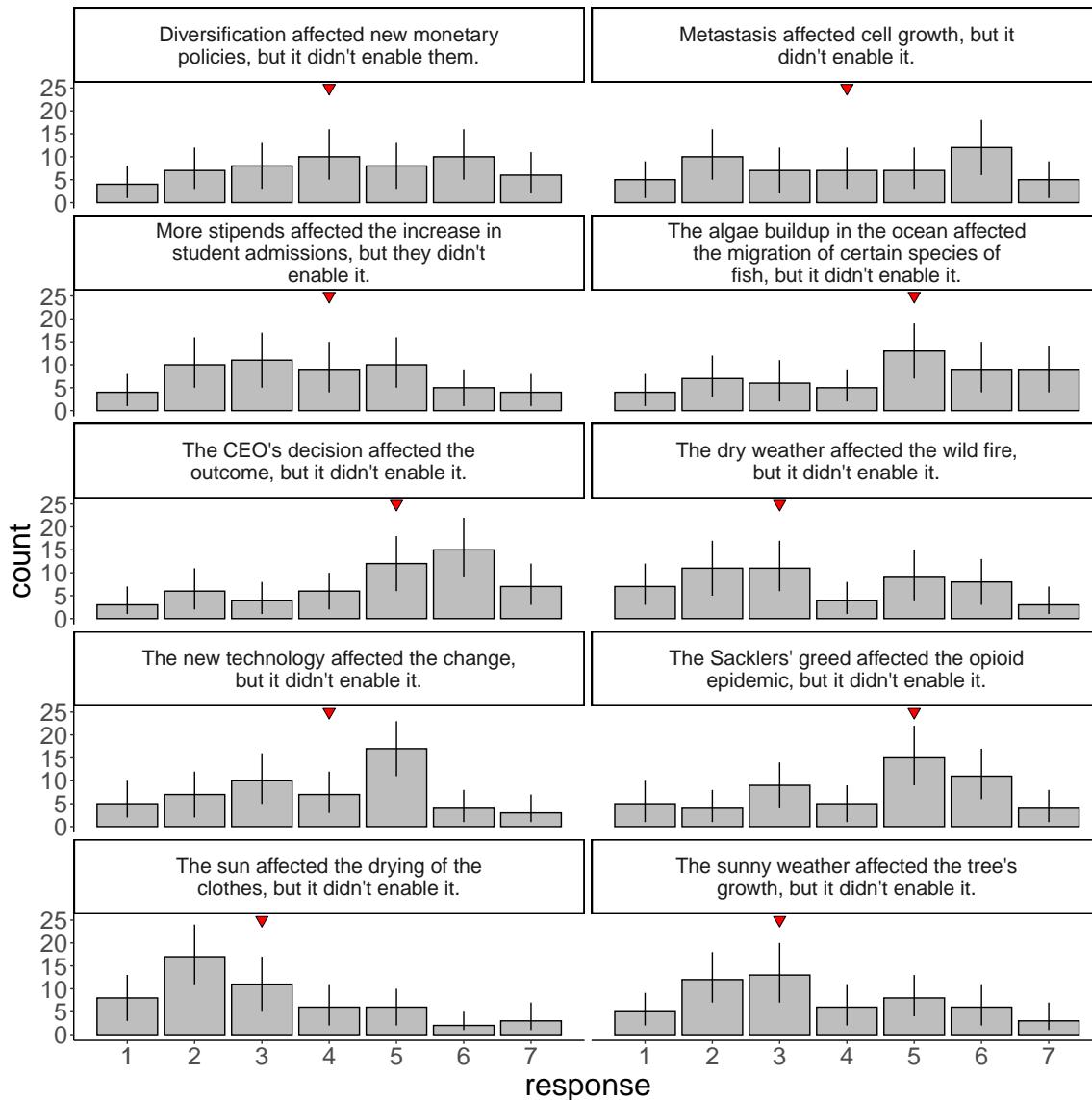


Figure B1. Histograms of participant responses in Experiment 1B for all items with the affected → enabled order. Note: 1 = “definitely not acceptable”, 4 = “unsure”, 7 = “definitely acceptable”. There is substantial variety in the distributions of participants’ responses, with some frames skewing toward the acceptable side of the scale and others skewing to the unacceptable side. Median rating is indicated by red triangle. Error bars represent bootstrapped 95% confidence intervals.

Appendix C
Experiment 1C Caused → Enabled Items

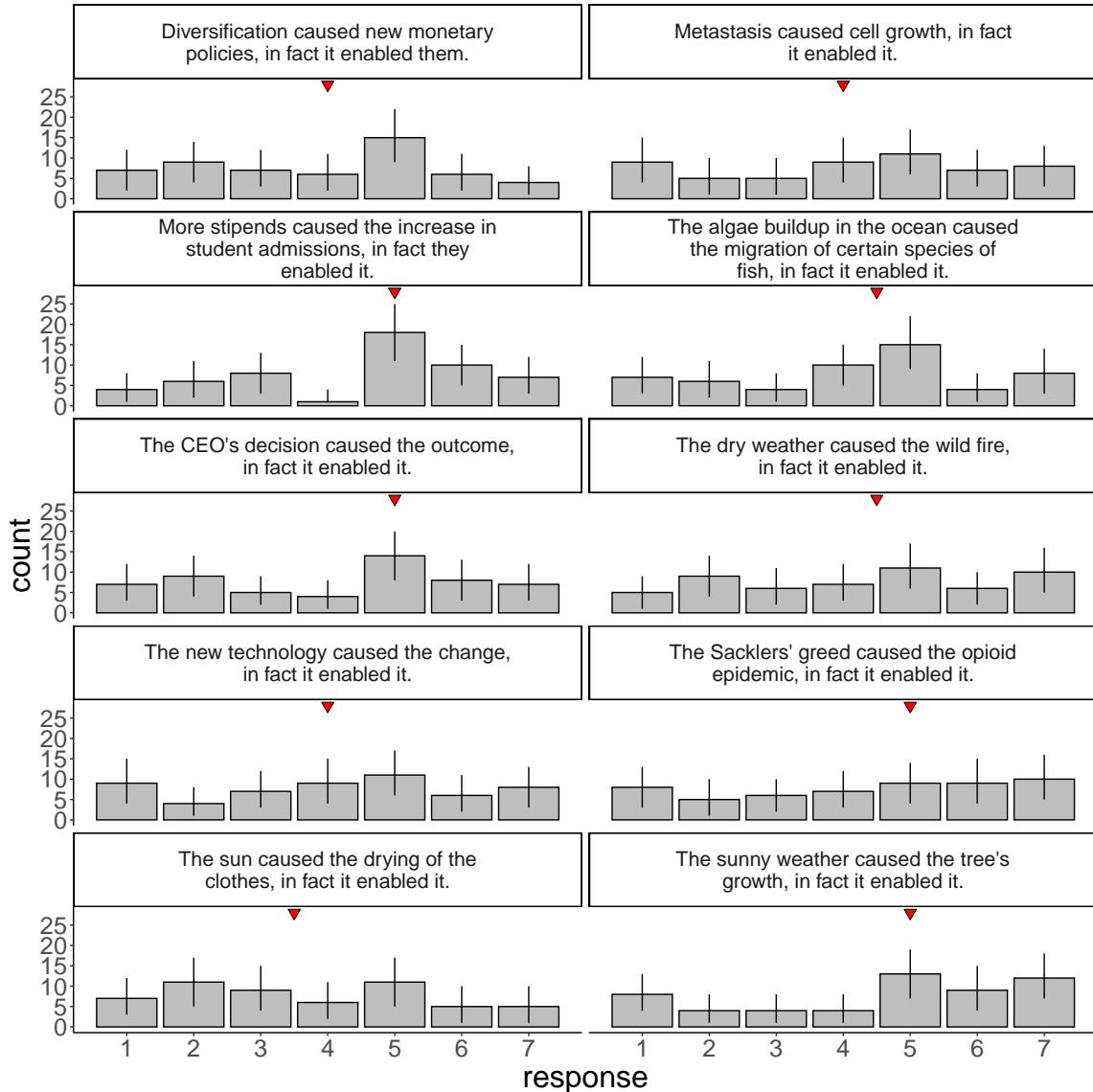


Figure C1. Histograms of participant responses in Experiment 1C for all items with the caused → enabled order. Note: 1 = “definitely not acceptable”, 4 = “unsure”, 7 = “definitely acceptable”. Again we see substantial variability within and between frames. Some response distributions appear more uniform, while others skew more to acceptability. Median rating is indicated by red triangle. Error bars represent bootstrapped 95% confidence intervals.

Appendix D
Scenario Schematics and Aspect Values

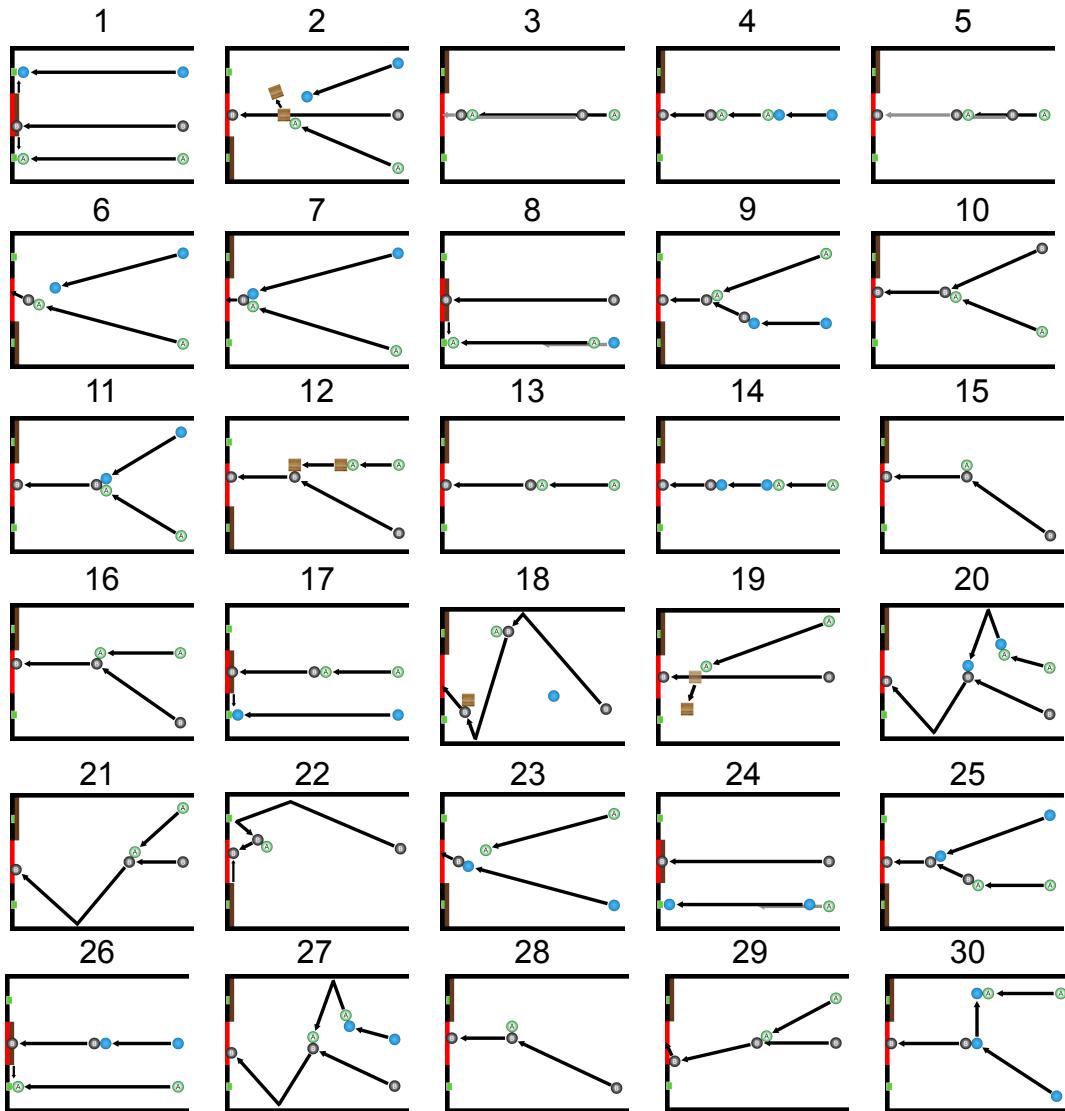


Figure D1. Trial schematics for all video scenarios in Experiment 2 and Experiment 3.

Table D1

Table of the aspect values for the scenarios from Experiment 2 and 3. Scenario numbers correspond to the numbers in Figure D1. Aspects are computed with noise value $\theta = 0.9$; the optimal value found for the full model in Experiment 2.

Scenario	Whether	How	Sufficient	Moving	Unique
1	0.0	0	0.0	1	0
2	0.32	0	1.0	1	0
3	0.0	1	0.0	1	1
4	0.333	1	0.0	1	1
5	0.037	1	0.041	1	1
6	0.0	1	1.0	1	1
7	0.0	1	1.0	1	0
8	0.101	0	1.0	1	0
9	0.618	1	0.0	1	0
10	0.785	1	0.795	1	1
11	0.905	1	0.039	1	0
12	0.997	1	0.001	1	0
13	1.0	1	1.0	1	1
14	1.0	1	1.0	1	0
15	0.999	1	0.999	0	1
16	0.999	1	0.999	1	1
17	1.0	1	0.0	1	1
18	0.314	1	0.036	0	0
19	1.0	0	1.0	1	0
20	0.957	1	0.048	1	0
21	0.22	1	0.241	1	1
22	1.0	1	1.0	0	1
23	0.0	0	0.0	1	0
24	0.0	0	0.0	1	0
25	1.0	1	0.39	1	0
26	1.0	0	0.0	1	0
27	0.972	1	0.037	1	1
28	0.81	1	0.791	0	1
29	0.08	1	0.059	1	1
30	0.0	0	0.0	1	0

Appendix E
Experiment 2: Trial Response Distributions

Table E1

Distribution of participant responses for all trials in Experiment 2. Scenario numbers correspond to schematics in Figure D1. Note: Some rows do not sum to 1 due to rounding.

Scenario	No Difference	Affected	Enabled	Caused
1	0.95	0.02	0.02	0.02
2	0.10	0.11	0.76	0.03
3	0.63	0.24	0.03	0.10
4	0.05	0.35	0.11	0.48
5	0.58	0.29	0.10	0.03
6	0.02	0.06	0.10	0.82
7	0.06	0.44	0.19	0.31
8	0.21	0.05	0.73	0.02
9	0.03	0.31	0.26	0.40
10	0.08	0.16	0.11	0.65
11	0.03	0.48	0.19	0.29
12	0.02	0.26	0.53	0.19
13	0.00	0.02	0.06	0.92
14	0.02	0.21	0.19	0.58
15	0.02	0.37	0.37	0.24
16	0.00	0.19	0.15	0.66
17	0.02	0.13	0.11	0.74
18	0.10	0.66	0.18	0.06
19	0.03	0.03	0.89	0.05
20	0.10	0.35	0.29	0.26
21	0.18	0.58	0.06	0.18
22	0.03	0.24	0.48	0.24
23	0.98	0.00	0.02	0.00
24	0.97	0.00	0.02	0.02
25	0.03	0.50	0.29	0.18
26	0.16	0.05	0.77	0.02
27	0.00	0.47	0.11	0.42
28	0.08	0.35	0.35	0.21
29	0.39	0.50	0.00	0.11
30	0.98	0.00	0.00	0.02

Appendix F
No Semantics and No Pragmatics: Top Regression Fit

Table F1

Fixed effects of the top performing No Semantics and No Pragmatics Model. The Estimate column gives the posterior mean for the given term, while the Estimate Error gives the posterior variance. The columns CI Lower Bound and CI Upper Bound give the lower and upper bounds of the 95% credible interval for the given term.

Term	Estimate	Estimate Error	CI Lower Bound	CI Upper Bound
Threshold No Difference Affected	2.42	0.64	1.16	3.67
Threshold Affected Enabled	3.74	0.64	2.47	4.98
Threshold Enabled Caused	4.76	0.64	3.48	6.01
whether	1.58	0.42	0.79	2.43
how	1.46	0.46	0.58	2.39
sufficient	1.54	0.37	0.82	2.29
moving	1.12	0.48	0.16	2.07
unique	0.16	0.39	-0.59	0.93

Appendix G
Experiment 3: Trial Pairings and Participant Summary Responses

Table G1

Video pairings for all trials in Experiment 3. The headers indicate trial expression. Video numbers refer to the scenarios in Figure D1. Means less than 0.5 indicate overall ratings favoring video 1, while means greater than 0.5 indicate overall ratings favoring video 2.

Trial	Video 1	Video 2	Mean	SD
Caused				
1	10	17	0.50	0.30
2	4	9	0.55	0.28
3	10	16	0.54	0.23
4	2	14	0.67	0.30
5	10	20	0.35	0.29
6	15	16	0.68	0.24
7	6	30	0.13	0.23
8	13	19	0.19	0.22
9	13	25	0.22	0.26
Enabled				
10	2	19	0.58	0.21
11	15	22	0.50	0.24
12	12	22	0.44	0.27
13	1	22	0.79	0.25
14	12	29	0.25	0.27
15	11	26	0.70	0.29
16	13	19	0.78	0.27
17	8	24	0.17	0.24
18	4	26	0.73	0.28
Affected				
19	17	28	0.69	0.28
20	15	16	0.44	0.24
21	5	7	0.69	0.25
22	14	21	0.48	0.28
23	25	26	0.42	0.30
24	7	23	0.21	0.27
25	18	24	0.21	0.27
26	18	23	0.19	0.23
27	18	30	0.14	0.20
Made No Difference				
28	3	29	0.28	0.27
29	23	30	0.49	0.23
30	5	29	0.29	0.25
31	3	12	0.20	0.29
32	3	14	0.20	0.29
33	5	7	0.28	0.28
34	8	24	0.72	0.31
35	1	21	0.23	0.29
36	1	27	0.18	0.28