Quantitative causal selection patterns in token causation

Adam Morris*  Jonathan Phillips  Tobias Gerstenberg
Harvard University  Harvard University  Stanford University
Fiery Cushman  Harvard University

Abstract

When many events contributed to an outcome, people consistently judge some more causal than others, based in part on the prior probabilities of those events. For instance, when a tree bursts into flames, people judge the lightning strike more of a cause than the presence of oxygen in the air – in part because oxygen is so common, and lightning strikes are so rare. These effects, which play a major role in several prominent theories of token causation, have largely been studied through qualitative manipulations of the prior probabilities. Yet, there is good reason to think that people’s causal judgments are on a continuum – and relatively little is known about how these judgments vary quantitatively as the prior probabilities change. In this paper, we measure people’s causal judgment across parametric manipulations of the prior probabilities of antecedent events. Our experiments replicate previous qualitative findings, and also reveal several novel patterns that are not well-described by existing theories.

I. Introduction

When something happens, people commonly ask: What caused that? Making causal judgments is often deceptively easy. We naturally conclude that the lightning strike caused the forest fire; the last-minute goal caused the sports team’s victory; or the scandal caused the political candidate’s defeat. These kinds of causal judgments are also very important, structuring how we understand and interact with our environments.

Yet, there are certain things that people are reluctant to label causes of an event, even though the event clearly depended on them. For instance, it is less natural to conclude that the forest fire was caused by the presence of oxygen, or by the lack of rain, or by the arsonist’s birth, even though all of these events were necessary for the fire to happen.

This phenomenon is widespread: When an outcome occurs, people consistently judge certain seemingly-relevant events more causal than others (Cheng & Novick, 1991; Halpern & Hitchcock, 2015; Hart & Honoré, 1959/1985; Icard, Kominsky, & Knobe, 2017). This phenomenon is sometimes labeled “specific” causation (Sloman, Barbey, & Hotaling, 2009; Walsh & Sloman, 2011), or “causal selection” (Hesslow, 1988; Hilton & Slugoski, 1986; McGill & Tenbrunsel, 2000). Understanding how and why people do this has been a central focus of philosophical and psychological inquiry for decades. (In this paper, we focus exclusively on token causation, where people judge the causes of a singular outcome, rather than type causation, where people judge the causes of an outcome in general.)

When attempting to understand why some causes stand out while others recede, a key fact is that this distinction is not dichotomous, but rather graded (Danks, 2017). Rather than judge one event entirely a cause and the others entirely not a cause, people appear to judge
causation on some kind of continuum. For instance, people might say that the arsonist was most causal; the lack of rain somewhat causal; and the arsonist’s birth the least causal. Although gradation was commonly assumed in many models of type causation (Cheng, 1997), the graded nature of token causation has only more recently attracted attention (Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017; Halpern & Hitchcock, 2015; Holyoak, Lee, & Lu, 2010; Icard et al., 2017; Stephan & Waldmann, 2016).

Despite its importance, however, our empirical understanding of gradation in token causal judgment is limited. Many studies have investigated how qualitative shifts in the parameters of causal systems (e.g., the prior probability of the antecedent events) produce qualitative shifts in causal judgment (see Icard et al. 2017 for an overview), but fewer have mapped the quantitative relationship between those parameters and the resulting judgments.¹ Here, we contribute to closing this empirical gap. We present people with two basic causal systems, systematically manipulate the prior probabilities of the antecedent events in those systems, and elicit people’s causal judgments. We find that, even in the most basic systems, the quantitative form of people’s judgments can be complex, and is incompletely described by existing theories. This result highlights the need for future research on the quantitative form of token causal judgments.

i. Qualitative manipulations of token causation

Several qualitative manipulations are known to affect which events people consider causal. We do not attempt an exhaustive review; rather, we follow Icard et al. (2017) and highlight certain robust ways in which a key feature of a deterministic causal system — the prior probabilities of the candidate causes — affects people’s judgment.

Conjunctive systems We first consider conjunctive causal systems, where multiple antecedent events were counterfactually necessary for an outcome — e.g., there was a fire which only occurred because of both oxygen and a lit match. In these systems, classic normative analyses of causation would, roughly, label the necessary events as equally causal (Lewis, 1973). Yet, people consistently judge an antecedent event more causal if it is rarer — i.e., if, a priori, it was less likely to occur.

This effect has been demonstrated in two ways. First, when asked to select which of several necessary events was the cause of an outcome within some naturally occurring series of events, people tend to pick the rarer event. For instance, people say the match, not the oxygen, caused the fire. They do so in part because oxygen is always present, and a lit match is rarer (Hart & Honoré, 1959/1985). This basic finding has been replicated in many scenarios (Hilton & Slugoski, 1986; Kahneman & Miller, 1986; Kahneman & Tversky, 1982).

Second, researchers have directly manipulated the prior probability of antecedent events, and found that making an event rarer causes people to assign it more causality. For instance, Icard et al. (2017) gave people vignettes like the following:

Professor Smith works at a large university. At this university, in order to get new computers from the university, faculty like Prof. Smith must send an application to two administrative committees, the IT committee and the department budget committee. Prof. Smith will be able to get her new computers if the IT committee AND the department budget committee approve her application.

Icard et al. manipulated the prior probability of one of the causes by telling participants that the budget committee either “almost always” or “almost never” approves applications. Then, participants were told that both of Prof. Smith’s applications were approved, and she received a computer. Participants were asked.

¹For a notable exception, see (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012, 2014, 2015).
the extent to which they agreed that the focal cause – the budget committee’s approval – caused Prof. Smith to receive the computer. Icard et al. found that people agreed substantially more when the budget committee almost never approves applications. This result, along with many similar results, demonstrates that rarer events are often considered more causal. Following others, we refer to this effect as “abnormal inflation”.

A less thoroughly explored effect is that, in conjunctive systems, people also judge an antecedent event less causal if the other necessary events are rare (Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015). For instance, if the budget committee almost never approves applications, then people judge the IT committee less causal; the budget committee “supersedes” the other potential causes. This is known as the supersession effect (Icard et al., 2017; Kominsky et al., 2015).

Disjunctive systems In disjunctive causal systems, where multiple events are each sufficient to produce an outcome, these effects change. Here, people appear to judge events more causal if they are less rare – an effect labeled “abnormal deflation” (Icard et al., 2017). For instance, if either committee can unilaterally approve Prof. Smith’s application, then the stingier committee is judged less causal. Moreover, the supersession effect appears to be eliminated in disjunctive structures (Icard et al., 2017; Kominsky et al., 2015). These findings, however, are less well-studied.

Summary In sum, we’ve highlighted four causal selection effects in token causal judgment: abnormal inflation and supersession in conjunctive systems, and abnormal deflation and no supersession in disjunctive systems (Icard et al., 2017). There are two features of these results worth noting. First, these causal selection effects occur even when people already know the underlying causal structure. Many models capture how people infer unobserved causal structures from sparse data (Cheng, 1997; Griffiths & Tenenbaum, 2005; Holyoak et al., 2010; Stephan & Waldmann, 2016), but these models are not designed to explain why people consider some variables more causal than others in the situations we study here. Our focus is on causal selection effects when people already know the causal structure.

Second, these causal selection effects have typically been studied with qualitative manipulations of the rarity of the antecedent events. Relatively little is known about how token causal judgments vary quantitatively with the probability of antecedent events. This is our point of departure.

ii. Our experimental design

To examine the quantitative nature of token causal judgment, we consider a deterministic causal system in which two binary variables combine, either conjunctively (Experiment 1) or disjunctively (Experiment 2), to produce a binary outcome (Figure 1A). In two experiments, we systematically manipulated the prior probabilities of the antecedent variables taking on value 1, and asked people the extent to which they believed one of the two variables was the cause of the outcome.

As discussed above, previous research has established that the prior probabilities of the antecedent variables matter for causal judgment. However, prior work has largely restricted itself to qualitative manipulations—e.g. comparing cases where an antecedent event is rarely present to cases where it is almost always present. Here, we map the prior probability of each antecedent variable (denoted \( C_1 \) and \( C_2 \)) to ten values: 0.1, 0.2, ..., 1. By independently manipulating \( \text{Prob}(C_1 = 1) \) and \( \text{Prob}(C_2 = 1) \) across the ten values, we produce one hundred different parameterizations of the causal system. We present these hundred parameterizations to people, and ask them to what extent they agree with the statement that \( C_1 \) caused the outcome.

To preview our results: We largely replicate the first three qualitative findings described above (abnormal inflation and supersession in
conjunctive structures, abnormal deflation in disjunctive structures), but we find a “reverse” form of supersession in disjunctive structures. Moreover, our experiments reveal nonlinear effects which, to our knowledge, are not well captured by existing theories.

The data and analysis code for both experiments can be found at https://github.com/adammorris/ causality.

II. Experiment 1: Conjunctive scenario

In Experiment 1, we presented participants with the following vignette and an associated image (Figure 1B):

A person, Joe, is playing a casino game where he reaches his hand into two boxes and blindly draws a ball from each box. He wins a dollar if and only if he gets a green ball from the left box and a blue ball from the right box.

Joe closes his eyes, reaches in, and chooses a green ball from the first box and a blue ball from the second box. So Joe wins a dollar.

Please tell us much you agree or disagree with this statement: Joe’s first choice (where he chose a green ball from the first box) caused him to win the dollar.

The images presented the two boxes, illustrating the percentage of green balls in the left box and blue balls in the right box. By manipulating these images, we manipulated the prior probability that Joe draws a green ball or a blue ball. Since we ask people to rate the causality of drawing the green ball, we will refer to green as the “focal” variable (labeled focal) and blue as the “alternate” variable (labeled alternate).

We divide the continuous space of probability into ten values: \( \text{Prob}(\text{focal} = 1) = 0.1 \), \( \text{Prob}(\text{focal} = 1) = 0.2 \), ..., \( \text{Prob}(\text{focal} = 1) = 1 \).

For each observation of the system, we randomly selected a prior probability for focal and alternate (the two are independent), and showed the participant the corresponding image. We then told the participant that Joe wins the dollar, and asked them to rate the extent to which they agree that Joe drawing the green ball caused him to win the dollar. Participants responded on a scale from 1 (“totally disagree”) to 9 (“totally agree”).

i. Methods

All participants were recruited through Amazon Mechanical Turk. They gave informed consent, and the study was approved by Harvard’s Committee on the Use of Human Subjects.

We employ a within-between subjects design. Each participant was given the vignette five times, each time with a randomly chosen probability setting. (We ensured that a participant did not see the same probability setting twice.) We ran \( N = 999 \) subjects, and collected a total of 4964 ratings. (No subjects were excluded, but some subjects did not complete all five ratings.) There are 100 different probability settings (10 for focal crossed with 10 for alternate), which gives an average of about 50 individual ratings per probability setting.

ii. Theoretical predictions

In addition to the qualitative predictions described above (abnormal inflation and supersession in the conjunctive structure), several contemporary theories of token causation make more precise quantitative predictions in our task. In our treatment of these models, we sometimes make additional assumptions to adapt them to the present task. Our intent is not to critically evaluate the models in their exact original setting, but rather to examine the extent to which plausible instantiations of them can capture quantitative variation in the situations we study.

\(^2\)We excluded zero because the outcome that participants actually observe is that Joe wins, which cannot happen if drawing either ball has probability zero.
Halpern and Hitchcock (2015) One influential theory comes from Halpern and Hitchcock (2015). Halpern and Hitchcock argue that when people evaluate counterfactuals by considering alternative possible worlds, they prefer to imagine worlds that are more normal, where unlikely things haven’t occurred (Kahneman & Miller, 1986; Kahneman & Tversky, 1982). For instance, people tend to imagine a world in which an arsonist didn’t light a match rather than a world in which there were no oxygen in the air. Or, in our case, if it’s rare to draw a green ball but common to draw a blue ball, they would tend to imagine an alternative world in which a green ball was not drawn.

Halpern and Hitchcock (2015) argue that this observation can explain the dependence of causal judgment on the prior probabilities of the variables. To determine whether a variable \( C \) was counterfactually necessary for an outcome, we consider an alternate world — a “witness” world — in which the variable is off (and perhaps some other variables have taken on different values; for details, see Halpern and Hitchcock (2015); Halpern and Pearl (2005)). Call this world \( w_C \). Halpern and Hitchcock propose that people will only judge a variable a cause of an outcome if (a) there exists a witness world in which \( C \) was necessary for the outcome, and (b) if that witness world—\( w_C \)—was more normal than the actual world. Moreover, if two variables \( C_1, C_2 \) meet these criteria, then \( C_1 \) will be considered more of a cause if \( w_{C_1} \) is more normal than \( w_{C_2} \).

Halpern and Hitchcock’s model is not committed to a particular normality ordering over possible witness worlds (for a detailed discussion, see Halpern and Hitchcock (2013)). In our task, however, it is likely that people’s normality ordering would be based on the probability of drawing the two balls; for instance, if the probability of drawing green is greater than 0.5, then a world in which it is drawn is more normal than a world in which it is not drawn. By making this assumption, we can derive a predicted response profile in our experiment (Figure 2A). If the probability of drawing a green ball is greater than 0.5, then the green ball should never be considered a cause, because \( w_{\text{focal}} \)—the world in which green is off, which is the sole witness world for green (the focal variable)—is less likely than the actual world. (In the actual world, a green ball was drawn; in \( w_{\text{focal}} \), a green ball was not drawn. If \( \text{Prob}(\text{focal} = 1) > .5 \), then, all else being equal, \( w_{\text{focal}} \) is less likely than the actual world.) This result is depicted in the bottom part of Figure 2A. If the probability of drawing
green is less than 0.5 and also less than the probability of drawing blue, then green is assigned a high causal rating, because \( w_{\text{focal}} \) is more likely than \( w_{\text{alternate}} \). This is depicted in the top part of Figure 2A. Finally, if the probability of drawing green is less than 0.5 but greater than the probability of drawing blue, then green is causal but less causal than blue; it is assigned an intermediate rating.

Formally, then, the degree to which focal was a token cause of Joe’s win is:

\[
TC_{HH}(\text{focal} \rightarrow \text{dollar}) =
\begin{cases}
0 & \text{if } \text{Prob}(\text{focal}) > .5 \\
x_{\text{high}} & \text{if } \text{Prob}(\text{focal}) < \text{Prob}(\text{alternate}) \leq .5 \\
x_{\text{mid}} & \text{if } \text{Prob}(\text{alternate}) < \text{Prob}(\text{focal}) \leq .5
\end{cases}
\]

where \( 0 < x_{\text{mid}} < x_{\text{high}} \leq 1 \).

Halpern and Hitchcock’s model is ordinal, and so it cannot assign precise numbers for \( x_{\text{high}} \) and \( x_{\text{mid}} \). Here, for simplicity, we assume that \( x_{\text{high}} = 1 \) and \( x_{\text{mid}} = \frac{1}{2} \). Critically though, our interest is in the overall shape of the model predictions, not the model’s precise numerical ratings.

**Icard et al. (2017)** Another theory we consider comes from Icard et al. (2017). Icard et al. offer a variation of the counterfactual approach. They propose that, when evaluating causality, people sample a counterfactual world with probability proportional to how likely that world is. For instance, to evaluate whether focal = 1 was the cause of Joe’s win, they sample a world in which focal = 0 (he didn’t draw a green ball) in proportion to \( \text{Prob}(\neg \text{focal}) \), and a world in which focal = 1 (he did draw a green ball) in proportion to \( \text{Prob}(\text{focal}) \). Then, crucially, people evaluate a different counterfactual depending on what they sampled. If they sampled a world in which focal = 0, they evaluate whether focal was necessary (holding all else about the actual world fixed); if they sampled a world in which focal = 1, they evaluate whether focal is, in general, sufficient (allowing other variables to vary).

According to Icard et al., while the probability of sampling a counterfactual world with focal = 0 is monotonically related to \( \text{Prob}(\text{focal} = 0) \), the two are not necessarily equal; there may be other factors that influence which worlds come to mind. This feature makes it difficult to derive precise quantitative predictions in our experiment. Nonetheless, if we make the simplifying assumption that the probability of sampling counterfactual worlds comes directly from the prior probability of green, then we can derive a response profile:

\[
TC_{\text{Icard}}(\text{focal} \rightarrow \text{dollar}) =
\begin{cases}
\text{Prob}(\neg \text{focal}) \cdot \text{Prob}(\text{focal} \text{ was necessary})+ \\
\text{Prob}(\text{focal}) \cdot \text{Prob}(\text{focal} \text{ would be sufficient})
\end{cases}
\]

which in our simple structure reduces to:

\[
TC_{\text{Icard}}(\text{focal} \rightarrow \text{dollar}) =
1 - \text{Prob}(\text{focal}) \cdot \text{Prob}(\neg \text{alternate})
\]

The predictions of this simplified version of Icard et al.’s model are shown in Figure 2B.

**Process theories** An alternate class of “process” theories eschew counterfactuals entirely, and explains token causation judgments in terms of a physical process connecting antecedent events to the outcome. On these views, the match was a cause of the fire because, e.g., it transferred some conserved physical properties to the fire (Dowe, 2000; Walsh & Sloman, 2011; Wolff, 2007; Wolff, Barbey, & Hausknecht, 2010). These theories predict that causal judgment will not meaningfully vary with the probabilities of antecedent events in a causal system like ours; we do not include them in our analysis.

**Classic causal strength measures** Finally, we consider three classic “causal strength” measures, some of which were designed primarily to capture other flavors of causal judgments (e.g. type causation). Nonetheless, these measures make plausible quantitative predictions in the causal systems we test, and so we include them in our analysis.

The first, denoted SP, assigns causality to
a variable to the extent to which observing the variable raises the probability of the outcome: \( TC_{SP}(C \rightarrow E, u_i) = \text{Prob}(E \mid C) - \text{Prob}(E) \) (Spellman, 1997). In our simple setup, this becomes: \( TC_{SP}(\text{focal} \rightarrow \text{dollar}) = \text{Prob}(\text{alternate}) \cdot \text{Prob}(\neg \text{focal}) \) (Figure 2C).

The second is similar, but requires causes to raise the probabilities of their outcomes, relative to a state where the cause was absent. This model, denoted Delta-P, takes the form: \( TC_{Delta-P}(C \rightarrow E, u_i) = \text{Prob}(E \mid C) - \text{Prob}(E \mid \neg C) \) (Jenkins & Ward, 1965; Perales & Shanks, 2007). In our case, this becomes: \( TC_{Delta-P}(\text{focal} \rightarrow \text{dollar}) = \text{Prob}(\text{alternate}) \) (Figure 2D).

The third builds on the second, but normalizes the rating by the probability that the event is absent when the cause is absent. This model, called Power-PC, takes the form: \( TC_{PPC}(C \rightarrow E, u_i) = \frac{\text{Prob}(E \mid C) - \text{Prob}(E \mid \neg C)}{\text{Prob}(E \mid \neg C)} \) (Cheng, 1997). In our case, this is identical to Delta-P: \( TC_{PPC}(\text{focal} \rightarrow \text{dollar}) = \text{Prob}(\text{alternate}) \) (Figure 2D).\(^3\)

There are other notable models of token causal judgment that use Bayesian reasoning to infer underlying causal structures from sparse data (Griffiths & Tenenbaum, 2005; Holyoak et al., 2010; Stephan & Waldmann, 2016). These models tend to explain causal selection effects in terms of inferred asymmetries in the underlying "causal power" of variables (e.g. when some variables have higher weights in a noisy-OR structure). In our experiments, however, people already know the underlying causal structures, and the structures have no inherent causal asymmetries; the only asymmetries are in the prior probabilities of the variables. Hence, we do not include these models in our analysis.

\(^3\)There is an alternate form of Power-PC which was designed specifically to assign causal attribution ratings when both the cause and outcome are known to have occurred (Cheng & Novick, 2005; it is the original formula divided by \( \text{Prob}(\text{outcome} \mid \text{cause}) \)). In both of our experiments, this alternate formula predicts a constant rating of 1 across all conditions. Hence, we do not include it in our analysis.

iii. Results

People’s average ratings are shown in Figure 2E, scaled from their original range of 1 to 9 to a range of 0 to 1. People clearly exhibit abnormal inflation; on average, they rate drawing green (the focal variable) as more causal when drawing green was less likely. They also clearly exhibit supersession; on average, they rate drawing green as less causal when drawing blue (the alternate variable) was less likely.

To demonstrate both of these effects statistically, we estimated a linear mixed effects model, regressing the causal ratings of green on the prior probabilities of drawing green and blue, with random intercepts and slopes for each subject. On average, people indeed rated green as more causal when its probability was lower (\( \beta = -0.02, SE = 0.002, t(959.9) = -11.8, p < .001 \)) and when the probability of drawing blue was higher (\( \beta = 0.03, SE = 0.001, t(949.5) = 18.9, p < .001 \)).

More interesting, however, are the nonlinear patterns. We highlight four such patterns, all revolving around what happens when drawing one of the balls is certain. First, people exhibit a large jump in abnormal inflation from \( \text{Prob}(\text{focal}) = 0.9 \) to \( \text{Prob}(\text{focal}) = 1 \) (Fig. 3A). In other words, when an event went from very likely to certain, this change caused an outsized drop in people’s causal rating of that event. To test this statistically, we regressed people’s causal ratings on \( \text{Prob}(\text{focal}) \), excluding trials where \( \text{Prob}(\text{focal}) = 1 \). We then used this regression model to predict people’s causal ratings for \( \text{Prob}(\text{focal}) = 1 \), and test whether their actual ratings when the focal variable is certain fall outside the 95% confidence interval of the regressions’s predictions. Indeed, people’s average causal rating when \( \text{Prob}(\text{focal}) = 1 \) is lower than what would be expected based on the ratings in the other conditions (predicted confidence interval was [0.41, 0.44]; average rating was 0.27).

Second, people exhibit a large jump in supersession from \( \text{Prob}(\text{alternate}) = 0.9 \) to \( \text{Prob}(\text{alternate}) = 1 \) (Fig. 3A). In other words, when the alternate event (drawing blue) went
Figure 2: Predictions and results from Experiment 1. (A-D) Predictions from the five models we analyze. The x-axis indicates the probability of drawing a green ball; line color indicates the probability of drawing a blue ball; and the y-axis indicates focal’s predicted causal rating. (E) Empirical results; y-axis indicates people’s agreement that drawing a green ball caused the outcome. Error bars depict +/-1 SEM.
from very likely to certain, this change caused an outsized increase in people’s causal rating of the focal event (drawing green). Using the same statistical approach as above, we find that the average rating when drawing blue is certain was higher than what would be expected based on the ratings in the other conditions (predicted confidence interval was [0.53, 0.55]; average rating was 0.70).

Third, people do not exhibit abnormal inflation when \( \text{Prob}(\text{alternate}) = 1 \)—except for a large decrease when \( \text{focal} \) also becomes certain (dark line in Fig. 3C). To test this, we regressed people’s ratings on \( \text{Prob}(\text{focal}) \), including only trials in which \( \text{Prob}(\text{alternate}) = 1 \) and excluding trials in which \( \text{Prob}(\text{focal}) = 1 \). Here, the probability of drawing green had no effect on people’s judgment (\( p = .77, BF_{\text{null}} = 9.0 \)). Moreover, there was a significant interaction: People showed more abnormal inflation when \( \text{Prob}(\text{alternate}) < 1 \), compared to when \( \text{Prob}(\text{alternate}) = 1 \) (\( \beta = .01, t = 2.1, p = .03 \)).

Fourth, there is an analogous effect for supersession; people do not exhibit supersession when \( \text{Prob}(\text{focal}) = 1 \) (except for perhaps a small effect when \( \text{alternate} \) is also certain; dark line in Fig. 3D). We tested this in the same way as above, regressing people’s ratings on \( \text{Prob}(\text{alternate}) \), including only trials in which \( \text{Prob}(\text{focal}) = 1 \) and excluding trials in which \( \text{Prob}(\text{alternate}) = 1 \). When the focal variable was certain, the probability of the alternate variable had no effect on people’s judgment (\( p = .31, BF_{\text{null}} = 5.7 \)). Moreover, there was a significant interaction: People showed more supersession when \( \text{Prob}(\text{focal}) < 1 \), compared to when \( \text{Prob}(\text{focal}) = 1 \) (\( \beta = -.01, t = -2.1, p = .04 \)).

**Item-level correlations between model and data** To compare people’s responses to the model predictions, we computed the correlation between the empirical ratings and each model’s predictions, across settings of \( \text{Prob}(\text{focal}) \) and \( \text{Prob}(\text{alternate}) \) (Figure 4). Icard et al.’s model and SP showed the highest correlation with people’s ratings (\( r’s = .73 \)), and did not differ significantly from each other (\( p = .9 \)). Icard et al.’s model was significantly more correlated with people’s ratings than the Halpern and Hitchcock model or the Delta-P/Power-PC model (\( p’s < .05 \)). SP was significantly more correlated with people’s ratings than the Halpern and Hitchcock model (\( p < .05 \)), and marginally more correlated than the Delta-P/Power-PC model (\( p = .08 \)).

The Icard model and SP both scored well by predicting the linear trends of abnormal inflation and supersession. However, neither correctly characterized the overall nonlinear patterns in people’s ratings. In particular, none captured the effects of certainty described above.

**IV. Discussion**

We characterized people’s quantitative causal ratings across the free parameters of a conjunctive causal system. As expected, people exhibit both abnormal inflation, where their causal rating of the focal variable (green) increased as it became rarer, and supersession, where their causal rating of the focal variable decreased as the alternate (blue) became rarer.

Moreover, people showed clear nonlinear patterns in judgment when the candidate causes were certain to occur. Specifically, they showed an unexpectedly strong drop in their causal rating of the focal variable when it was certain to occur, and an unexpectedly strong increase when the alternate was certain to occur. Moreover, the effect of abnormal inflation disappeared when the alternate was certain to occur, and the effect of supersession disappeared when the focal variable was certain to occur. These patterns are not well-described by extant models.

**III. Experiment 2: Disjunctive scenario**

Experiment 2 was identical to Experiment 1, except the causal structure was disjunctive instead of conjunctive (Fig. 5A). We used the same vignette, but told participants that Joe won a dollar if either a green ball was drawn
**Figure 3:** Breaking down the patterns in Experiment 1. (A) People show a larger abnormal inflation effect from \( \text{Prob}(\text{focal}) = 0.9 \) to \( \text{Prob}(\text{focal}) = 1 \). (B) People show a larger supersession effect from \( \text{Prob}(\text{alternate}) = 0.9 \) to \( \text{Prob}(\text{alternate}) = 1 \). (C) People show no abnormal inflation effect when the alternate variable is certain, except when the focal variable is also certain. (D) People show no supersession effect when the focal variable is certain (except perhaps a small effect when the alternate variable is also certain.) Error bars and shaded intervals are SEM. Regression lines in (C) and (D) are estimated with Loess regression to visualize nonlinearities around certainty.
from the left box, or a blue ball was drawn from the right box (Fig. 5B).

On each trial, participants were told that Joe drew both a green and blue ball, and won the dollar. As before, we manipulated the prior probabilities of drawing green (Prob(focal = 1)) and blue (Prob(alternate = 1)), and asked people the extent to which they agreed that Joe’s drawing the green ball caused him to win the dollar.

i. Methods

All participants were recruited through Amazon Mechanical Turk, and gave informed consent. We ran $N = 1013$ subjects, each of which gave 5 ratings (for a total of 5039 ratings and an average of about 50 ratings per probability setting).

ii. Theoretical predictions

In the disjunctive structure, the qualitative predictions from past research were (a) that participants would show abnormal deflation (making an event rarer would make it seem less causal), and (b) that participants would not show supersession (making the alternate event rarer would not affect causal judgment; see Icard et al. (2017); Kominsky et al. (2015)). For more precise, quantitative predictions, we turned to the same models as in Experiment 1.

Halpern & Hitchcock  Recall that in Halpern and Hitchcock’s model, a variable is considered a cause if the world we must imagine to judge whether it was necessary—its “witness” world—is more normal than the actual world. (As before, we assume that, in our task, normality ordering is determined by the probabilities of drawing green and blue.) In Experiment 2, to judge whether drawing a green ball was necessary for winning the dollar, we must consider a possible world — call it $w_{\text{focal}}$ — in which neither a green nor a blue ball was drawn, because it is only when alternate = 0 that focal could have been necessary. (For more details of
Quantitative patterns of causal selection in token causation

Figure 5: Causal structure (A) and vignette (B) used in Experiment 2.

how their model treats overdetermined cases like Experiment 2, see Halpern and Hitchcock (2015). Hence, the model predicts that the causal status of focal will depend on whether that world — \( w_{\text{focal}} \) — is more or less normal than the actual world.

To determine the precise predictions, there are three cases to consider. First, consider the case where both \( \text{Prob}(\text{focal}) \) and \( \text{Prob}(\text{alternate}) \) are less than 0.5. In this case, \( w_{\text{focal}} \) (where neither ball is drawn) is surely more normal than the actual world (where both balls are drawn), and the model predicts that focal will be judged causal. Moreover, as long as \( \text{Prob}(\text{focal}) < 0.5 \) and \( \text{Prob}(\text{alternate}) < 0.5 \), the extent to which focal is judged a cause is not predicted to vary with either probability; the only other candidate cause, alternate, has exactly the same witness world, and so the relative causality of focal will not vary as the normality of that witness world changes. Hence, in this case, the model assigns focal a constant causal rating, which we label \( x_{\text{high}} \).

Similarly, if \( \text{Prob}(\text{focal}) \) and \( \text{Prob}(\text{alternate}) \) are both greater than 0.5, then \( w_{\text{focal}} \) is less normal than the actual world, and the model predicts that focal will be judged noncausal. Hence, in this case, the model assigns focal a causal rating of zero.

The third and hardest case is when one of the two variables is likely, and the other is rare—that is, when \( \text{Prob}(\text{focal}) < 0.5 \) and \( \text{Prob}(\text{alternate}) > 0.5 \), or when \( \text{Prob}(\text{focal}) > 0.5 \) and \( \text{Prob}(\text{alternate}) < 0.5 \). Here, there are at least two ways of instantiating Halpern and Hitchcock’s model. From one perspective, the actual world and \( w_{\text{focal}} \) are, in this case, incomparable; one variable takes on its more typical value, but another takes on its less typical value, and the model offers no machinery for integrating those two facts. This is the view adopted in the original paper (Halpern & Hitchcock, 2015). When the actual world cannot be compared to the witness world, the model assigns a causal rating of zero—and hence focal will be judged noncausal.

But it is plausible that, in our experiment, people would rank the normality of a world according to the total probability of its variables. On this view, focal will only be considered causal if the total probability of the actual world (where both balls are drawn) — \( \text{Prob}(\text{focal}) \cdot \text{Prob}(\text{alternate}) \) — is less than the total probability of the witness world (where neither ball is drawn) — \( \text{Prob}(\neg \text{focal}) \cdot \text{Prob}(\neg \text{alternate}) \). This condition reduces to \( \text{Prob}(\text{focal}) + \text{Prob}(\text{alternate}) < 1 \).

This leaves us with two potential response profiles, HH1 and HH2:

\[
T_{C_{\text{HH1}}}(\text{focal} \rightarrow \text{dollar}) =
\begin{cases} 
  x_{\text{high}} & \text{if } \text{Prob}(\text{focal}) < 0.5 \text{ and } \text{Prob}(\text{alternate}) < 0.5 \\
  0 & \text{otherwise}
\end{cases}
\]
Quantitative patterns of causal selection in token causation

\[ TC_{HH2}(\text{focal} \rightarrow \text{dollar}) = \]
\[
\begin{cases} 
\ x_{\text{high}} & \text{if } \text{Prob}(\text{focal}) + \text{Prob}(\text{alternate}) < 1 \\
\ 0 & \text{otherwise}
\end{cases}
\]

Below, we consider both variations of the model. Their predictions are depicted in Figure 6A and B, assuming \( x_{\text{high}} = 1 \).

**Icard et al.** Recall that, according to the model of Icard et al. (2017):

\[
TC_{Icard}(\text{focal} \rightarrow \text{dollar}) = \]
\[
\text{Prob}(\neg \text{focal}) \cdot \text{Prob}(\text{focal} \text{ was necessary}) + \]
\[
\text{Prob}(\text{focal}) \cdot \text{Prob}(\text{focal} \text{ would be sufficient})
\]

In the structure of Experiment 2, this equation reduces to:

\[
TC_{Icard}(\text{focal} \rightarrow \text{dollar}) = \text{Prob}(\text{focal})
\]

The Icard model’s predictions are shown in Fig. 6C.

**Classic causal strength measures** In the disjunctive structure of Experiment 2, the three causal strength measures considered above reduce to (Figure 6D-F):

\[
TC_{SP}(\text{focal} \rightarrow \text{dollar}) = \]
\[
\text{Prob}(\neg \text{focal}) \cdot \text{Prob}(\neg \text{alternate})
\]
\[
TC_{DP}(\text{focal} \rightarrow \text{dollar}) = \text{Prob}(\neg \text{alternate})
\]
\[
TC_{Power-PC}(\text{focal} \rightarrow \text{dollar}) = 1
\]

**iii. Results**

People’s average ratings are shown in Figure 6G. As predicted by previous qualitative work (Icard et al., 2017), people exhibited abnormal deflation; they rated focal as less causal when it was rarer (\( \beta = .01, SE = .002, t = 8.8, p < .00001 \)). However, contrary to previous work (Icard et al., 2017; Kominsky et al., 2015), there was also a “reverse” supersession effect: People rated focal as less causal when alternate was more common (\( \beta = -.005, SE = .001, t = -3.7, p = .0003 \)).

We dissect these effects further in Figure 7. The abnormal deflation effect is largely linear; unlike with abnormal inflation in conjunctive structures, here there is no large jump from \( \text{Prob}(\text{focal}) = 0.9 \) to \( \text{Prob}(\text{focal}) = 1 \) (Fig. 7A). The reverse supersession effect, however, does show a nonlinearity: It is strongest when drawing blue becomes certain (Fig. 7B). We tested this statistically in the same way as in Experiment 1, and found that people’s ratings when \( \text{Prob}(\text{alternate}) = 1 \) were significantly lower than expected based on a linear model of the other conditions (predicted confidence interval was [.67, .70]; average rating was .64). Notably, the reverse supersession effect, though small, was still significant after excluding the condition when drawing blue was certain (\( \beta = -.004, SE = .002, t = -2.4, p = .02 \)).

There were two other nonlinear patterns that mirrored the patterns in Experiment 1. Recall that, in Experiment 1, people stopped exhibiting abnormal inflation when drawing blue was certain, and stopped exhibiting supersession when drawing green was certain (with certain exceptions discussed above). In Experiment 2, we find an analogous pattern: People stop exhibiting abnormal deflation when drawing blue is certain (Fig. 7C), and stop exhibiting reverse supersession when drawing green is certain (Fig. 7D). These visual observations were confirmed statistically. When \( \text{Prob}(\text{alternate}) = 1 \), there was no effect of \( \text{Prob}(\text{focal}) \) on people’s ratings (\( p = .75, BF_{null} = 9.5 \)); moreover, the effect of \( \text{Prob}(\text{focal}) \) was significantly stronger when \( \text{Prob}(\text{alternate}) < 1 \) compared to when \( \text{Prob}(\text{alternate}) = 1 \) (interaction \( \beta = -.01, p = .02 \)). Similarly, when \( \text{Prob}(\text{focal}) = 1 \), there was no effect of \( \text{Prob}(\text{alternate}) \) on people’s ratings (\( p = .55, BF_{null} = 8.5 \)); moreover, the effect of \( \text{Prob}(\text{alternate}) \) was stronger when \( \text{Prob}(\text{focal}) < 1 \) compared to when \( \text{Prob}(\text{focal}) = 1 \), although in this case the interaction was marginal (\( \beta = .01, p = .095 \)).
Figure 6: Predictions and results from Experiment 2. (A-F) Predictions from the models we analyze. The x-axis indicates the probability of drawing a green ball; line color indicates the probability of drawing a blue ball; and the y-axis indicates focal’s predicted causal rating. (E) Empirical results; y-axis indicates people’s agreement that drawing a green ball caused the outcome. Error bars are SEM.
Figure 7: Breaking down the patterns in Experiment 2. (A) People exhibit a linear effect of abnormal deflation. (B) People show a larger reverse supersession effect from \( \text{Prob(\text{alternate})} = 0.9 \) to \( \text{Prob(\text{alternate})} = 1 \), although the effect is still significant after excluding the trials where drawing blue is certain. (C) People don't show abnormal deflation when drawing blue is certain. (D) People don’t show reverse supersession when drawing green is certain. Error bars and shaded intervals are SEM; regression lines in (C) and (D) are estimated with standard sum-of-squares.
Model correlations  The correlations between the model predictions and people’s ratings across conditions are shown in Figure 8. The Icard model showed by far the best correlation with people’s ratings ($r = .64, p < .00001$); it was significantly more correlated with the empirical ratings than the next-best model ($p = .0005$). As in Experiment 1, however, it did not capture all the major patterns in people’s responses — most notably, it did not capture the reverse supersession effect in Fig. 7B, or either of the nonlinear patterns in Fig. 7C and 7D.

iv. Discussion

We characterized people’s quantitative causal ratings across the free parameters of a disjunctive causal system. As predicted by the Icard model (Icard et al., 2017), people exhibited abnormal deflation, where their causal rating of the focal variable decreased as it became rarer. Surprisingly, they also exhibited a “reverse” supersession effect, where their causal rating of the focal variable increased as the alternate variable became rarer.

Moreover, people’s nonlinear patterns mirrored those in Experiment 1: They stopped exhibiting abnormal deflation when the alternate variable was certain to occur, and stopped exhibiting reverse supersession when the focal variable was certain to occur. These nonlinearities are not well-described by extant models.

IV. General Discussion

We presented people with two basic causal systems, and elicited people’s causal judgments across parametric manipulations of two key parameters in those systems: the prior probability of the focal event (whose causal status was being rated), and the the prior probability of an alternate event (that was either necessary or sufficient for the outcome). These experiments replicated three previously established qualitative effects — abnormal inflation and supersession in the conjunctive structure, abnormal deflation in the disjunctive structure — and revealed a novel effect of reverse supersession in the disjunctive structure. Moreover, people exhibited a constellation of nonlinear patterns that were strikingly consistent across the two experiments. We discuss these results in turn.

Reverse supersession  The original supersession effect is that, in conjunctive structures, a focal event is rated less causal when an alternate event becomes rarer; the alternate event “supersedes” the focal event (Kominsky et al., 2015). This effect was only supposed to apply, however, to conjunctive structures; two previous papers predicted and found that there would be no supersession in disjunctive structures (Icard et al., 2017; Kominsky et al., 2015).

In contrast to those papers, we do find a kind of supersession effect in the disjunctive structure of Experiment 2, however, the effect is reversed: the focal event is rated more causal when the alternate event becomes rarer (or, equivalently, the focal event is rated less causal when the alternate event becomes more common). Given the logic of supersession, this reversal is unsurprising. In disjunctive structures, events become causally preferred when they are more common (the “abnormal deflation” effect). It makes sense, then, that if the alternate event were to supersede the focal event, it would do so when it was common, not rare.

What is surprising is that, in contrast to previous work, the alternate event supersedes the focal event at all in the disjunctive structure. There are three ways to reconcile this result with the null findings in previous papers (Icard et al., 2017; Kominsky et al., 2015). First, in our data, the supersession effect in disjunctive structures was indeed much weaker than the corresponding supersession effect in conjunctive structures. This is in line with the findings from previous work. Second, previous papers relied on qualitative manipulations of the probability of the alternate event; this may have made it difficult to detect the subtle effect reported here. Third, previous papers used different types of vignettes; it is possible that supersession in disjunctive structures is unique to the type of experimental manipulation used
Quantitative patterns of causal selection in token causation

Figure 8: Correlation between model predictions and empirical ratings for Experiment 2. Brackets show 95% confidence intervals for Pearson correlation coefficients. The correlation for Power-PC cannot be calculated because it predicts a constant response profile.
here. Future work should test for disjunctive supersession in a wider range of vignettes.

Nonlinear patterns around certainty Across most of the parameter space, the effects of prior probabilities on causal judgment appear, to a first approximation, to be linear and additive. That is, the prior probabilities of the focal and alternate events each have a linear influence on causal judgment, and do not interact much. The main exception is when the probability of either event approaches 1. When one of the events is certain to occur, people exhibit a constellation of departures from the linear, additive baseline.

First, three of the four primary effects – abnormal inflation and supersession in Experiment 1, reverse supersession in Experiment 2 – become very strong when the relevant events become certain. For instance, in Experiment 1, although the focal variable (drawing a green ball) is considered less and less causal as it becomes more common, the drop in causal rating from $\text{Prob}(\text{focal}) = 0.9$ to $\text{Prob}(\text{focal}) = 1$ is much larger than the drops between other conditions (See Fig 3A).

Discontinuities in causal judgment when an event is certain have been discussed before. Perhaps most prominently, Cheng and Novick (1991) argued that, when an event is certain to occur, people treat it as an “enabling condition”, not a cause (see also Einhorn and Hogarth (1986); Goldvarg and Johnson-Laird (2001); Mackie (1974); Walsh and Sloman (2011); Wolff et al. (2010)). The categorical distinction between cause and enabling condition may help explain why, in the conjunctive structure, people are particularly hesitant to label the focal event a cause when it was certain to occur. However, these “enabling condition” theories have difficulty explaining why people judged the certain events somewhat causal (instead of at zero causalilty), and why people judged the certain events to be highly causal in the disjunctive structure.

Another possibility is that these nonlinearities result from a domain-general feature of people’s probabilistic reasoning: People, in general, treat the jump from almost-certain to certain as more important than other equivalently sized jumps in probability (Kahneman & Tversky, 2013; Tversky & Kahneman, 1981). This finding, known as the “certainty effect”, could help explain why the patterns in causal judgment become more pronounced when an event becomes certain.4

Both of these perspectives, however, have trouble accounting for the second set of departures from the linear, additive baseline: People stop exhibiting all the effects described here (abnormal inflation/deflation, supersession) when the other event is certain to occur. Concretely, the prior probability of the focal event stops influencing people’s causal judgment when the alternate event is certain to occur, and the probability of the alternate event stops influence people’s judgment when the focal event is certain to occur. These results, which are consistent across both experiments should be further examined in future research.

Comparing extant models Out of the models analyzed here, the response profile predicted by Icard et al. (2017) was consistently the most similar to people’s ratings. This is perhaps not surprising, since Icard et al. built their model with some of our focal effects—i.e. abnormal selection and supersession—in mind. Still, it is satisfying that their model, which was designed principally to capture qualitative effects, roughly tracks the overall shape of people’s quantitative responses in our experiments. However, no model analyzed here captured all the nonlinear patterns in people’s judgments, highlighting the need for further theorizing.

Conclusion People’s causal judgments are influenced by the prior probabilities of the candidate causes. Here, we find that people show systematic patterns of judgment across parametric manipulations of those probabilities, in a way that is not fully characterized by existing theories.
theories. We hope this spurs further research into the quantitative nature of token causal judgment.

ACKNOWLEDGEMENTS

We thank Joshua Knobe, Thomas Icard, and the Moral Psychology Research Lab for their advice and assistance. AM, JP, and FC were supported by Grant N00014-14-1-0800 from the Office of Naval Research to FC. TG was supported by the Center for Brains, Minds & Machines (CBMM), which is funded by the National Science Foundation’s Science and Technology Center (Award CCF-1231216).

REFERENCES


doi: 10.1016/j.cognition.2017.01.010


