An interaction effect of norm violations on causal judgment (in press at *Cognition*)

Maureen Gill, Jonathan Kominsky, Thomas Icard, Joshua Knobe

June 4, 2022

Abstract

Existing research has shown that norm violations influence causal judgments, and a number of different models have been developed to explain these effects. One such model, the necessity/sufficiency model, predicts an interaction pattern in people's judgments. Specifically, it predicts that when people are judging the degree to which a particular factor is a cause, there should be an interaction between (a) the degree to which that factor violates a norm and (b) the degree to which another factor in the situation violates norms. A study of moral norms (N = 1000) and norms of proper functioning (N = 3000) revealed robust evidence for the predicted interaction effect. The implications of these patterns for existing theories of causal judgments is discussed.

1 Introduction

A department office keeps a collection of pens. Administrative assistants are allowed to take the pens, but faculty members are supposed to buy their own. One day, there are two pens left in the office collection. An administrative assistant takes one, and a faculty member takes the other. Now there is a problem: there are no pens left. When participants in an experiment receive this case, they tend to say that the faculty member caused the problem (Knobe and Fraser, 2008).

Now consider a slightly different scenario. This time, both administrative assistants and faculty members are allowed to take pens. An administrative assistant takes one pen, and a faculty member takes the other. Again, there is a problem: no pens left. In this latter case, participants are significantly less inclined to say that the faculty member caused the problem (Phillips et al., 2015).

Effects like this one seem to point to an impact of prescriptive norms on causal judgments. People's judgments about whether, and to what extent, an event caused some effect seem to differ depending on whether the event is seen as bad or wrong in

some way. This type of effect has been shown in numerous existing studies (Knobe and Fraser, 2008; Phillips et al., 2015; Kominsky et al., 2015; Icard et al., 2017).

A growing number of theories and models have emerged to explain and predict the influence of norm violations on causal judgment (Driver, 2008; Alicke et al., 2011; Samland and Waldmann, 2016). In particular, a growing literature has leveraged computational or formal modeling to predict the impact of scenario manipulations, such as the good/bad manipulation described above (Halpern and Hitchcock, 2015; Icard et al., 2017; Morris et al., 2021; Quillien, 2020; Blanchard and Schaffer, 2017). These models have the advantage of generating highly specific, testable predictions. For example, as applied to the pen scenario, these models are able to predict higher causal judgment of the professor's action when the action is not allowed vs. allowed.

In this paper, we derive and test a particular prediction from one model, the "necessity/sufficiency" model (Icard et al., 2017). It is known that in cases like the pen vignette where two events are needed for an effect, participants are more inclined to think that an event is causal if the event is regarded as bad than if it is regarded as good. The necessity/sufficiency model predicts that the magnitude of this effect, i.e., the size of the difference between the case in which the event is bad and the case in which it is good, should depend in a very specific way on whether *another* of the events is bad or good.

1.1 Abnormal inflation and supersession

A substantial literature has documented causal judgment in scenarios with a structure like the one in the pen vignette, that is, where two events are necessary for some effect to occur (but neither is alone sufficient to bring about the effect). The event that participants rate is referred to as the "focal event" and the other event is the "alternate event" (Morris et al., 2019). For example, in the pen vignette if participants are asked whether the professor's action caused the effect, then the professor's action is the focal event and the administrative assistant's action is the alternate event.

The literature to date has found that causal judgments of the focal event are affected by:

1. whether the focal event is good or bad
2. whether the alternate event is good or bad

Table 1 illustrates these effects by showing the full 2×2 for the pen vignette. The two effects documented in the existing literature are (1) the main effect of the goodness or badness of the focal event and (2) the main effect of the goodness or badness of the alternate event.

The main effect of the focal event's normative status (good/bad) on judgments of the focal event's causality is called "abnormal inflation": broadly, an event will be judged as more casual if it violates a prescriptive norm than if it does not (Knobe and

	Assistant ("alternate") bad	Assistant ("alternate") good
Professor ("focal") good	Professor allowed Assistant not allowed	Professor allowed Assistant allowed
Professor ("focal") bad	Professor not allowed Assistant not allowed	Professor not allowed Assistant allowed

Table 1: Possible manipulations of pen vignette

Fraser, 2008; Alicke, 1992). For example in the pen vignette, the professor is judged as more causal when her action is not allowed compared to a version of the vignette where her action is allowed (Phillips et al., 2015).

The normative status of the alternate event also affects judgments of the focal event's causality, an effect called "supersession." Specifically, the focal event is judged as less causal when the alternative event is bad than when the alternate event is good (Kominsky et al., 2015). For example, the professor would be judged less causal if the administrative assistant's action is prohibited than if the administrative assistant's action were allowed.

Numerous further studies have explored the effect of prescriptive norms on causal judgments. The effect has been shown in children as young as five when presented with a child-friendly version of the pen vignette (Samland et al., 2016). The effect also persists, for example, even when professors are not allowed to take pens but regularly do (Roxborough and Cumby, 2009). The effect applies not only to actions, but also to omissions (Henne et al., 2016), and not only to retrospective judgments (about what has already occurred), but also to prospective judgments (what will occur; Henne et al. 2021b). The effect also occurs in reverse: Kirfel et al. (2021) used both vignettes and visual stimuli to illustrate that participants are able to infer information about norms from causal statements. Finally, studies on judgment about controversial moral questions show that the extent to which an individual participant regards a certain action as morally wrong predicts the degree to which that participant will regard the action as causal: those participants who view abortion as wrong, for example, are more likely to judge a doctor performing an abortion as the cause of a further outcome (Cushman et al., 2008).

Though the examples presented so far only highlight moral norms, or rules of how people should behave in a particular situation, the same pattern of results also occurs for other kinds of prescriptive norms, such as how an artifact or natural kind ought to work. Violations of these "norms of proper functioning" have also shown inflation (Hitchcock and Knobe, 2009; Livengood et al., 2017) and supersession (Kominsky and Phillips, 2019). For example, participants presented with a vignette about a battery that would short circuit if two wires touched it at the same time judged a wire as more causal if it was not supposed to touch the battery (Hitchcock and Knobe, 2009). For ease of reference, throughout this paper, we will refer to an event as "bad" when

it violates a prescriptive norm of any kind and "good" when it does not, including both moral norms and norms of proper functioning.

1.2 Necessity/sufficiency model

As we noted above, there are numerous different theories within the existing literature that offer explanations of the impact of prescriptive norms on causal judgment (Alicke et al., 2011; Driver, 2008; Halpern and Hitchcock, 2015; Quillien, 2020; Samland and Waldmann, 2016). In this Introduction, we focus just on one of those theories: the necessity/sufficiency model (Icard et al., 2017). This theory generates a surprising prediction about the pattern of people's intuitions that we will be putting to the test in the present studies. In the General Discussion, we then consider a wider array of different theories and ask what implications the findings from the studies might have for each of them.

At the core of the necessity/sufficiency model are three key claims. First, there is the claim that people make causal judgments by sampling possibilities in a causal model. The necessity/sufficiency model applies this sampling idea to one specific problem, but the approach more generally has been applied to numerous problems both in computer science and in cognitive science, primarily as a means of approximating difficult probabilistic calculations (MacKay, 2003). Inspired by the success of sampling based approximations in engineering, similar models have been proposed as "algorithmic" or "process-level" accounts of human inductive reasoning (Sanborn and Chater, 2016; Icard, 2016).

According to the necessity/sufficiency model, causal judgment involves an iterative process of "sampling" possibilities from a probability distribution. (For example, the process might involve considering the possibility: "What if the professor had refrained from taking a pen?"). People then perform a kind of check regarding that one possibility. After they have considered a number of different possibilities, they arrive at a judgment based on the result of all the different checks they have performed. As the number of samples goes to infinity (and in fact, typically much faster) this judgment will converge toward a particular value. It is assumed that people typically take only a few samples, and their judgments will therefore be some approximation of that value.

The second key claim is about the distribution from which people are sampling. One obvious hypothesis would be that the probability of sampling a particular possibility is just equal to people's belief about the probability of that possibility. (For example, if people think that the probability of the professor taking a pen is .62, it might be thought that the probability of sampling that possibility should be .62.) The necessity/sufficiency model rejects that obvious hypothesis in favor of something a bit more complex. The claim is that prescriptive norms can influence the probability of sampling a particular possibility. Thus, if people think that the professor really ought to refrain from taking a pen, then that belief—a belief about what ought to

happen—will influence the probability of sampling that possibility. In general, people will have a higher probability of sampling a given possibility to the extent that it conforms to prescriptive norms. Although we will be concerned in particular with the application of this idea to questions in causal cognition, this broad idea has also been applied in many other domains of cognitive science (Bear and Knobe, 2017; Lieder et al., 2018; Phillips et al., 2019).

The third claim is that the possibility people sample determines whether they check for necessity or sufficiency. For example, suppose people are wondering whether the professor's act of taking a pen was the cause of the problem. To address this question, they would sample some action that the professor could have taken. They might end up sampling a possibility in which the focal event does not occur (i.e., a possibility in which the professor does not take the pen) or a possibility in which the focal event does occur (i.e., a possibility in which the professor does take the pen). This sampling process will then determine whether they check for necessity or for sufficiency.

If they sample a possibility in which the focal event does not occur, they check for necessity. In other words, they consider a possibility in which the focal event does not occur and then simulate forward to imagine whether the effect would not occur. By contrast, if they sample a possibility in which the focal event does occur, they check for sufficiency. That is, they consider a possibility in which the focal event does occur, then imagine some way that other events could be (e.g., whether the alternate event occurs) and simulate forward to see whether the effect still occurs.

Putting these three claims together, we get predictions about how prescriptive norms will shape the process of making causal judgments. Specifically, if the focal event is seen as violating a prescriptive norm, people will check more for necessity, whereas if the focal event is seen as conforming to a prescriptive norm, people will check more for sufficiency. Then, if people do check for sufficiency, the normative status of the alternate event will play a role. If the alternate event is seen as violating a prescriptive norm, people will tend to check for necessity by imagining possibilities in which the alternate event does not occur, whereas if the alternate event is seen as conforming to a prescriptive norm, people will tend to check for sufficiency by imagining possibilities in which the alternate event does occur.

Within existing work, the necessity/sufficiency model has usually been spelled out more formally. To do this, we will use $P(F = 1)$ for the probability in each iteration of the algorithm of sampling a possibility in which the focal event F occurs and $P(A = 1)$ for the probability in each iteration of sampling a possibility in which the alternate event A occurs. It is worth emphasizing that these probabilities describe how likely the individual is to sample such a possibility, rather than, e.g., describing a subjective judgment about probability. In each iteration in which the algorithm checks to see whether the focal event F is necessary for an effect E, there is a certain probability that it will arrive at a positive result. We write this probability $P^{\nu}_{F=0}(E = 0)$. Similarly, in each iteration in which the algorithm checks to see whether the focal

event F is sufficient for an effect E, there is a certain probability that it will arrive at a positive result. We write this probability $P^{\sigma}_{F=1}(E = 1)$. The necessity/sufficiency model can then be spelled out in terms of Algorithm 1:

Algorithm 1: Determine the causal strength of F on E (with K samples)

Initialize $N = 0$.
for $k \leq K$ **do**
  Sample a value $F^{(k)}$ from P.
  if $F^{(k)} = 0$ **then**
    Sample $E^{(k)}$ from $P^{\nu}_{F=0}$. Let $N = N + (1 - E^{(k)})$. "necessity"
  else if $F^{(k)} = 1$ **then**
    Sample $E^{(k)}$ from $P^{\sigma}_{F=1}$. Let $N = N + E^{(k)}$. "sufficiency"
return N/K

As $K \to \infty$, Algorithm 1 converges toward the value

$$P(F = 0)P^{\nu}_{F=0}(E = 0) + P(X = 1)P^{\sigma}_{F=1}(E = 1).$$

We will be treating this as our measure of causal strength. Now consider the application of Algorithm 1 to causal structures in which the focal cause F and the alternative cause A are individually necessary and jointly sufficient for effect E. In this specific structure, if one sets F to 0 and then simulates forward, E will always turn out to be 0. Similarly, if one sets F to 1 and then simulates forward, E will turn out to be 1 if and only if $A = 1$. Thus, in this specific causal structure, $P^{\nu}_{F=0}(E = 0)$ will be 1, while $P^{\sigma}_{F=1}(E = 1)$ will be $P(A = 1)$. The degree to which the focal event F will be seen as a cause of effect E should then be:

$$P(F = 0) + P(F = 1)P(A = 1)$$

In the present studies, we will be concerned only with this one causal structure, and we will therefore be focusing on the predictions derived from it.

1.3 Explanations of current predictions

In cases that have this structure, the necessity/sufficiency model predicts a main effect of the goodness or badness of the focal event ("inflation") and a main effect of the goodness or badness of the alternate event ("supersession"), but importantly, it also predicts an interaction of these two factors.

To get a better understanding of this interaction, we can visualize the model predictions. Figure 1 shows the model predictions for cases like the pen vignette, where two events are necessary to bring about an effect. While it is possible to visualize the model predictions in a fine continuous gradient, as seen in Figure 1, the present experiment does not test thousands of fine distinctions on a continuum from

good to bad, but merely a binary distinction between whether the focal event was good or bad and whether the alternate event was good or bad.

Color is used to show predicted causal judgment (red is more causal, blue is less causal). The y-axis shows the probability of sampling possibilities in which the focal event occurs; the x-axis shows the probability of sampling possibilities in which the alternate event occurs. Recall that a key claim of the model is that the probability of sampling an event occurring depends in part on the degree to which that event is seen as good or bad. Thus, the x- and y-axes can also be seen as representing the degree to which the alternate and focal events are seen as good.

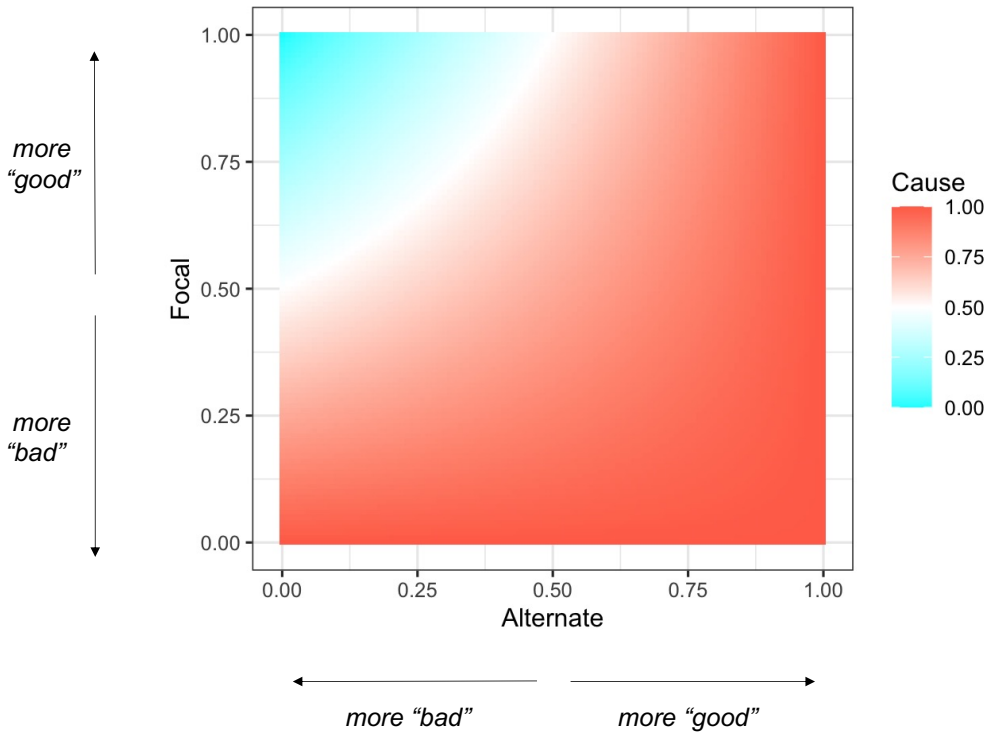

Figure 1: Model predictions for the necessity/sufficiency model, in the case of an unshielded collider structure with two variables that are individually necessary and jointly sufficient. Color shows predicted causal judgment. Model predictions are generated from the formula $P(F = 0) + P(F = 1)P(A = 1)$.

As can be seen in the figure, the necessity/sufficiency algorithm predicts two existing effects from the literature. First, there is "inflation," the main effect of the focal event on causal judgments. Moving along the y-axis, when the focal event is bad (i.e., lower value on the y-axis), it is more causal (i.e., a darker shade) than when the focal event is good (higher values on the y-axis). Second, there is "supersession," as

can be seen by following the x-axis. When the alternate is bad (left side of x-axis), the focal is judged as less causal than when the alternate is good (right side of y-axis). For a possible example of what the study results might look like according to the necessity/sufficiency model, see the Appendix.

Crucially, however, the figure also shows that in addition to the two main effects, there is a predicted interaction. This interaction effect can be characterized in two ways:

1. *Inflation increase.* The abnormal inflation effect is stronger when the alternate event is bad than when the alternate event is good. In Figure 1, this effect can be seen in that the difference between the top and the bottom is itself larger on the left than on the right.

2. *Supersession decrease.* The supersession effect is weaker when the focal event is bad than when the focal event is good. This effect can be seen in that the difference between the left and right is itself smaller on the bottom than on the top.

At an intuitive level, the necessity/sufficiency algorithm suggests a possible mechanistic story about how these effects (inflation, supersession, inflation increase/ supersession decrease) may occur.

The explanation of inflation is straightforward. On a given iteration, the algorithm will either check for necessity or sufficiency. The focal event is always necessary, but not always sufficient. Therefore, the degree to which the focal event is seen as causal is a function of how much the algorithm checks for necessity (as opposed to sufficiency). The algorithm checks for necessity more when it samples more possibilities in which the focal event does not happen, that is, in the case where the focal event is bad. Thus, the focal will be judged as more causal when it is bad than when it is good.

This same approach can be used to explain supersession. As mentioned, the normative status of the alternate event (good or bad) influences the computed sufficiency of the focal event. So when the alternate event is bad, the focal event will be seen as less sufficient and therefore less causal.

We thereby arrive at a mechanistic explanation for the interaction effect. The normative status of the alternate event (good or bad) only matters for the focal event's computed sufficiency, and not necessity. After all, the focal event is always necessary and so the value of the alternate event does not matter. Therefore, when the algorithm checks necessity more, the normative status of the alternate event exerts less of an effect. In other words, there is less supersession in the case that necessity is checked more, i.e., when the focal event is bad.

For a more formal analysis of the predicted interaction, see the Appendix. As the analysis shows, the necessity/sufficiency model predicts that if a given scenario shows both inflation and supersession, then it must also show an interaction. Moreover, the necessity-sufficiency model puts certain lower and upper bounds on the size of

this interaction. So given that one observes an inflation effect of a certain size and a supersession effect of a certain size, the necessity/sufficiency model says that there has to be an interaction of at least a certain size. At the same time, it should be noted that it would be compatible with the necessity/sufficiency model for the interaction effect to be substantially smaller than either the main effect of inflation or the main effect of supersession. For this reason, the studies reported here use sample sizes that would be sufficient to detect even a very small effect.

1.4 The present studies

The necessity/sufficiency model predicts an interesting possibility: an interaction effect of norm violations on causal judgment, as described above, characterized by "inflation increase" and "supersession decrease." Other than that this interaction is predicted by the model, we had no independent reason to believe that such an interaction effect exists. To empirically test this model prediction, we ran two experiments where we independently manipulated the norm status of the focal event and the alternate event.

Experiment 1 tests the predicted interaction effect in the moral domain. Experiment 2 aims to generalize the prediction beyond the moral domain, using stories about how organs and parts of a machine should work to determine whether the predicted interaction extends beyond moral norms.

2 Experiment 1

Participants read a scenario about two causes that are both needed to bring about some effect. Each cause either violated a moral norm or did not violate a moral norm.

We predicted that, when an event ("focal event") violates a moral norm, it should receive higher causal ratings than if it were normative: "inflation". Critically, we also pre-registered the prediction that there would be an interaction effect and that the interaction would show a specific pattern. We predicted that the magnitude of this increase will be greater when the other event ("alternate event") violates a norm: "inflation increase". Conversely, we predicted that the focal event should be seen as less causal when the alternate event violates a norm ("supersession"), but that the magnitude of this effect should decrease when the focal event violates a norm: "supersession decrease." Methods and predictions were pre-registered at: https://aspredicted.org/blind.php?x=be6ki3. Additionally, all data and analyses for Experiment 1 and Experiment 2 are available for download at: https://osf.io/t7us3.

2.1 Method

Participants. Participants were 1000 adults recruited from MTurk (463 men, 533 women, 4 other). Participation was restricted to MTurk workers in the United States who had completed at least 5000 past HITs with a minimum approval rating of 99%. An additional 383 participants were excluded on the basis of pre-registered exclusion criteria: failing the attention check question, failing to answer the critical causality question, or for using an IP address that another participant used. Three additional participants beyond the pre-registered sample size were accidentally collected due to experimenter error and are not included in the analyses.

Materials and Procedure. Participants were randomly assigned to read one of four versions of five different vignettes about Sam and Brook. Each vignette had two components that could be manipulated: the normative status of Sam's action and the normative status of Brook's action. The overall design was therefore a 2 (Sam "focal event": good vs. bad) × 2 (Brook "alternate event": good vs. bad) × 5 (vignette) design. Two of the vignettes were the "motion detector" and "computer" vignettes from Experiment 1 of Icard et al. (2017). Another was the "pen" vignette adapted from Knobe and Fraser (2008), modified to name the characters Brook and Sam. Additionally, we created two vignettes for the purpose of this experiment ("library" and "plumbing"). See Table 2 for one of the new vignettes and its four variants.

After reading a vignette, participants were asked to indicate the extent to which they agreed with the statement "Sam caused [the effect]" on a scale from 1 (strongly disagree) to 7 (strongly agree). Then, participants answered a check question about the prescriptive norms in the vignette, for example "Who was supposed to arrive at 9 am?" with the option to select "Sam," "Brook," "both," or "none". Participants were excluded for answering incorrectly. Finally, participants reported their gender, race/ethnicity, and SES (eight non-numeric options from "top of the ladder" to "bottom of the ladder"; also a "prefer not to say" option).

2.2 Results

Before proceeding with the primary analysis, we tested whether we successfully replicated the abnormal inflation and supersession effects with t-tests (not pre-registered). To test the abnormal inflation effect, we compared causal ratings (which were always of the focal event) in the two "focal good" conditions ($M = 2.991$, $SD = 1.765$) to the two "focal bad" conditions ($M = 5.498$, $SD = 1.414$), and found that the focal event was rated significantly more causal in the focal bad conditions, $t(999) = 24.686$, $p < .001$, successfully replicating the abnormal inflation effect. For the supersession effect, we compared causal ratings in the two "alternate bad" conditions ($M = 5.076$, $SD = 2.105$) to those in the "alternate good" conditions ($M = 3.831$, $SD = 1.911$), and found that the focal event was significantly less causal in the alternate bad conditions, $t(999) = 18.676$, $p < .001$, successfully replicating the causal superseding

Introduction. A new book by a best-selling author came out yesterday, and the local library was only able to secure two copies. Brook and Sam are at the library and are both looking forward to reading the new book.		
	Sam ("focal") bad	**Sam ("focal") good**
Brook ("alternate") good	Brook checked out one copy from the library. Sam realized he had considerable late fees and that he would not be allowed to take out the book. Sam looked around to make sure no one was watching and slipped a copy in his bag.	Sam checked out one copy from the library. Brook also checked out a copy.
Brook ("alternate") bad	Brook realized he had considerable late fees and that he would not be allowed to take out the book. Brook looked around to make sure no one was watching and slipped a copy in his bag. Sam also realized he had considerable late fees and that he would not be allowed to take out the book. Sam looked around to make sure no one was watching and slipped a copy in his bag.	Sam checked out one copy from the library. Brook realized he had considerable late fees and that he would not be allowed to take out the book. Brook looked around to make sure no one was watching and slipped a copy in his bag.
Effect. Some time later, a third library patron hoped to read the new book and wondered if there were any copies at the library. She looked at the shelf where the books would have been and realized that there were no copies left.		
Question. *Please indicate the degree to which you agree or disagree with the following statement:* "Sam caused the library to be without any copies of the new book."		

Table 2: Each of the four conditions of the "library" vignette, one of the five vignettes used in Experiment 1.

effect.

We then proceeded to the pre-registered analysis, constructing two linear mixed effects models, both treating the moral valence of each actor as a fixed factor, and treating vignette as a random factor with respect to intercept. We fit one model with the main effects of both fixed factors as well as their interaction, and a second model that removed the interaction term. An ANOVA comparison of the two models

revealed that the model with the interaction term better predicted causal judgments, $\chi^2(1, 1000) = 5.413$, $p = .019$.

As Figure 2a shows, this interaction is in the predicted direction. The abnormal inflation effect is larger when the alternate event is bad than when the alternate event is good; in other words, the supersession effect is larger when the focal event is good than when the focal event is bad.

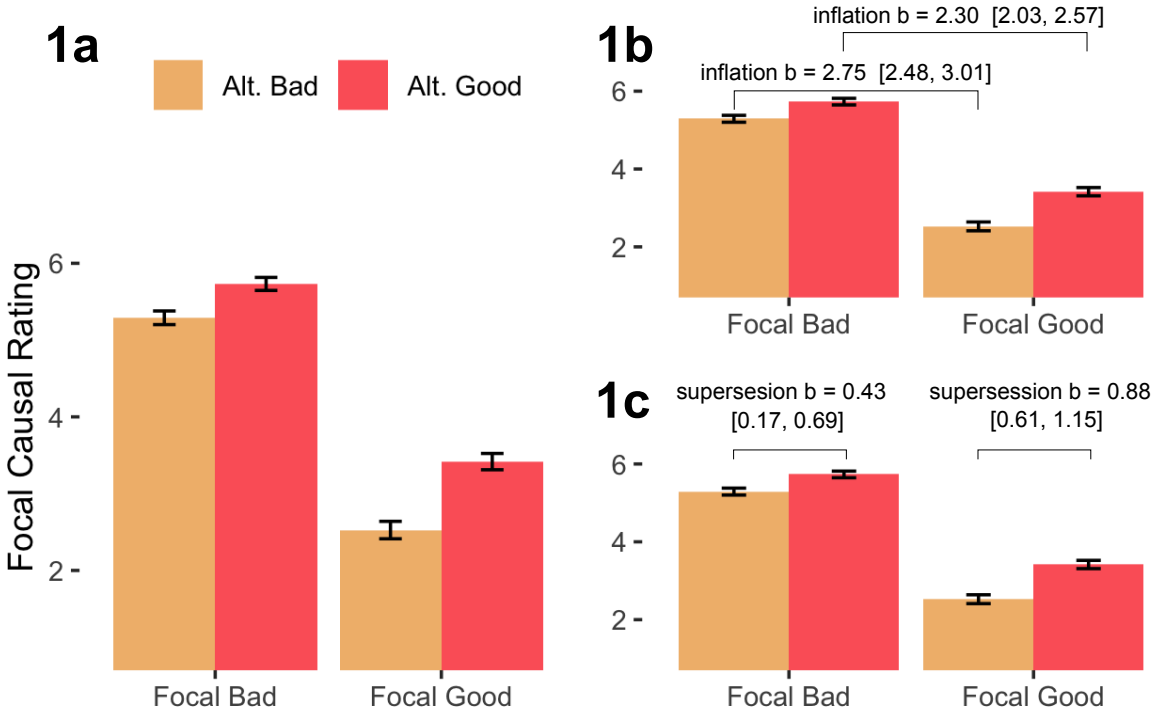

Figure 2: Results from Experiment 1. All three graphs depict the same results. 1a shows mean causal rating by focal event condition (bad vs. good) and alternate event condition (bad vs. good). 1b and 1c are the same graph as 1a and are reprinted to highlight the effect sizes. 1b compares the size of the abnormal inflation effect when the alternate event is good vs. when the alternate event is bad. 1c compares the size of the supersession effect when the focal event is good vs. when the focal event is bad.

To further explore this interaction, we ran pairwise comparisons using the "emmeans" package in R. The results showed that the effect size for inflation (2b) was indeed larger when the alternate was bad (2.75; 95% CI [2.48, 3.01]), $t(999) = 20.251$, $p < .001$, vs. good (2.30; 95% CI [2.03, 2.57]), $t(999) = 16.872$, $p < .001$, ("inflation increase"), albeit with slightly overlapping confidence intervals. Also, the effect size for supersession (2c) was smaller when the focal was bad (.43; 95% CI [.17, .69]), $t(999) = 3.252$, $p = .001$, vs. good (.88; 95% CI [0.61, 1.15]), $t(999) = 6.404$, $p < .001$ ("supersession decrease"), again with slightly overlapping confidence intervals.

2.3 Discussion

This first experiment replicated the findings from previous research that the focal agent is regarded as more causal when her actions violate a norm ("abnormal inflation") and less causal when an alternative agent's actions violate a norm ("supersession"). Critically, the results also showed the predicted interaction. The magnitude of the abnormal inflation effect depended on the alternative agent's actions, such that the difference between the focal agent being good and bad was greater when the alternate agent violated a moral norm ("inflation increase"). Accordingly, the magnitude of the supersession effect was smaller when the focal agent violated a moral norm ("supersession decrease").

3 Experiment 2

In Experiment 2, we explored whether the "inflation increase"/"supersession decrease" pattern based on moral norm violations in Experiment 1 would extend to violations of norms of proper functioning. Participants read a scenario about an object with two parts, where one part must be [working/not working] and the other part must be [working/not working] in order for some effect to occur.

Previous work has found that inflation and supersession occur in norms of proper functioning. We predicted that we would replicate these "inflation" and "supersession" effects found in previous studies of norms of proper functioning and additionally to find the "inflation increase"/"supersession decrease" found in Experiment 1. We predicted that, analogously to moral norms, when a part is not working properly, it should receive higher causal ratings ("inflation") and that the magnitude of this increase will be greater when the other part is not working: "inflation increase". Conversely, we predicted that we would replicate the finding that the "focal" part should be seen as less causal when the "alternate" part violates a norm ("supersession"), but that the magnitude of this effect should decrease when the focal event violates a norm: "supersession decrease." As with Experiment 1, methods and predictions were pre-registered at: https://aspredicted.org/GSK_FWN.

3.1 Method

Participants. Participants were 3000 adults recruited from MTurk (720 men, 2253 women, 35 other/prefer not to say). Participation was restricted to MTurk workers in the United States who had completed at least 5000 past HITs with a minimum approval rating of 97%. An additional 2060 participants were excluded for using an IP address that another participant used and/or failing the attention check question.

Materials and procedure. Participants were randomly assigned to read one of four versions of two different vignettes ("alien" vignette and "battery" vignette; see Table

3 for an example). Each vignette had two components that could be manipulated: whether one part was working properly and whether the other part was working properly. The overall design was therefore a 2 ("focal": good vs. bad) × 2 ("alternate": good vs. bad) × 2 (vignette: alien vs. battery) design.

After reading a vignette, participants were asked to indicate the extent to which they agreed with the statement "The buzzing sound occurred because the red wire touched the battery" ("battery" vignette condition) or "The alien turned blue because of the puzzer" ("alien" vignette condition) on on a scale from 1 (strongly disagree) to 7 (strongly agree). Then, participants answered a check question about the prescriptive norms in the vignette, for example in the battery vignette "Which wire is supposed to touch the battery?" with option "black wire" vs. "red wire." For the alien vignette, the check question was "Which organ was working normally?" with options for "puzzer" and "denizer."

Participants were excluded for answering incorrectly. Finally, participants reported their gender, rated how much attention they paid to the task on a 7 point scale from "almost no attention" to "my complete attention," and answered three open-ended feedback questions (e.g., "What did you think this survey was about?").

3.2 Results

Before proceeding with the primary analysis, we tested whether we successfully replicated the abnormal inflation and supersession effects with t-tests (not pre-registered). To test the abnormal inflation effect, we compared causal ratings (of the focal event) in the two "focal good" conditions ($M = 5.157$, $SD = 0.490$) to the two "focal bad" conditions ($M = 3.858$, $SD = 0.0433$), and found that the focal event was rated significantly more causal in the focal bad conditions, $t(2999) = 19.861$, $p < .001$, successfully replicating the abnormal inflation effect. For the supersession effect, we compared causal ratings in the two "alternate bad" conditions ($M = 3.831$, $SD = 1.971$) to those in the "alternate good" conditions ($M = 5.076$, $SD = 1.657$), and found that the focal event was significantly less causal in the alternate bad conditions, $t(2999) = 18.676$, $p < .001$, successfully replicating the causal superseding effect.

We then proceeded to the pre-registered analysis, constructing two linear mixed effects models, both treating the moral valence of each actor as a fixed factor, and treating vignette as a random factor with respect to intercept. We fit one model with the main effects of both fixed factors as well as their interaction, and a second model that removed the interaction term. An ANOVA comparison of the two models revealed that the model with the interaction term better predicted causal judgments, $\chi^2(1, 3000) = 20.835$, $p < .001$.

As Figure 3a shows, this interaction is in the predicted direction. The abnormal inflation effect is larger when the alternate event is bad than when the alternate event is good and the supersession effect is larger when the focal event is good than when the focal event is bad.

<table>
<tr><td colspan="3">Introduction. A machine is set up in such a way that it will make a buzzing sound if both the black wire and the red wire touch the battery at the same time. The machine will not make a buzzing sound if just one of these wires touches the battery.</td></tr>
<tr><td></td><td>Alternative (black) good</td><td>Alternate (black) bad</td></tr>
<tr><td>Focal (red) good</td><td>When the machine is switched on, the red wire and the black wire are supposed to touch the battery. The switch is supposed to put the red wire in contact with the battery and the black wire in contact with the battery. So, there will be a buzzing sound only if the black wire is put in the right place and the red wire is put in the right place.</td><td>When the machine is switched on, the red wire is supposed to touch the battery, while the black wire is supposed to be in some other part of the machine that does not touch the battery. The switch is supposed to put the red wire in contact with the battery and put the black wire in a place not touching the battery. So, there will be a buzzing sound only if the red wire is put in the right place and the black wire is put in the wrong place.</td></tr>
<tr><td>Focal (red) bad</td><td>When the machine is switched on, the black wire is supposed to touch the battery, while the red wire is supposed to be in some other part of the machine that does not touch the battery. The switch is supposed to put the black wire in contact with the battery and put the red wire in a place not touching the battery. So, there will be a buzzing sound only if the black wire is put in the right place and the red wire is put in the wrong place.</td><td>When the machine is switched on, the black wire and the red wire are supposed to be in some other part of the machine that does not touch the battery. The switch is supposed to put the black wire in a place not touching the battery and the red wire in a place not touching the battery. So, there will be a buzzing sound only if the black wire is in the wrong place and the red wire is in the wrong place.</td></tr>
<tr><td colspan="3">Effect. One day, when the switch is turned on, the switch puts the red wire on the battery and puts the black wire on the battery. There is a buzzing sound.</td></tr>
<tr><td colspan="3">Question. Please indicate the degree to which you agree or disagree with the following statement:
“The buzzing sound occurred because the red wire touched the battery.”</td></tr>
</table>

Table 3: Each of the four conditions of the “battery” vignette, one of the two vignettes used in Experiment 2.

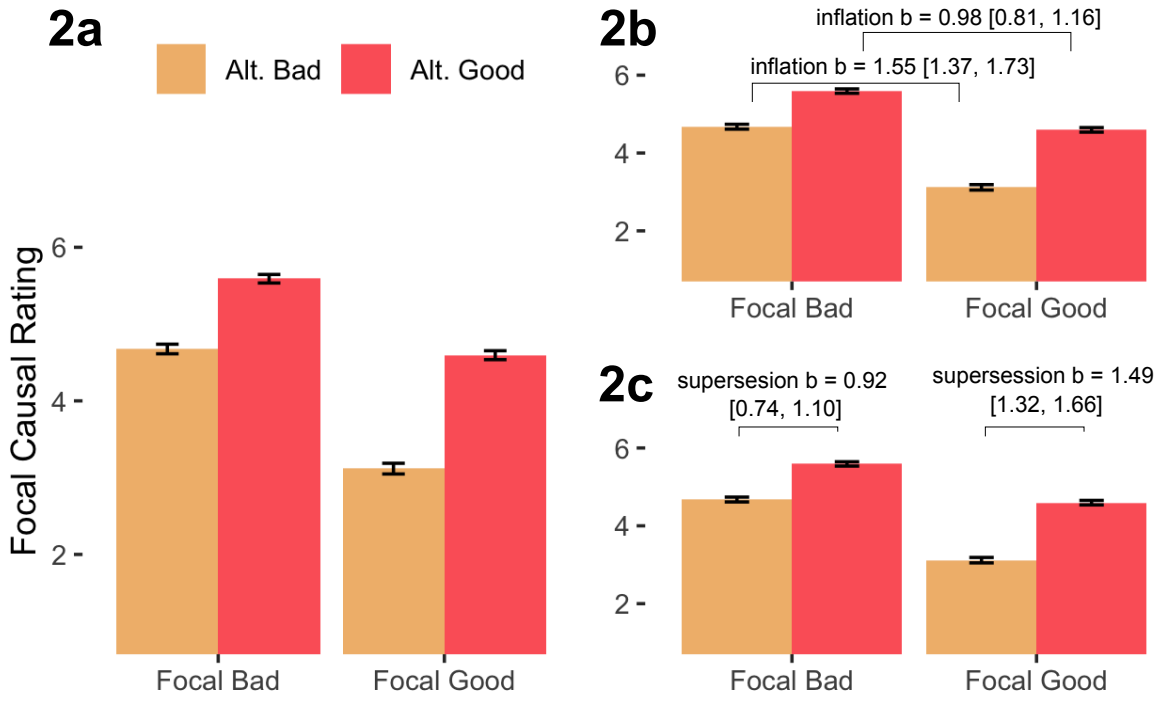

Figure 3: Results from Experiment 2. All three graphs depict the same results. 2a shows mean causal rating by focal condition (bad vs. good) and alternate condition (bad vs. good). 2b and 2c are the same graph as 1a and are reprinted to highlight the effect sizes. 2b compares the size of the abnormal inflation effect when the alternate event is good vs. when the alternate event is bad. 2c compares the size of the supersession effect when the focal event is good vs. when the focal event is bad.

As with Experiment 1, we ran pairwise comparisons to further explore this interaction. The results showed that the effect size for inflation (3b) was indeed larger when the alternate was bad (1.55; 95% CI [1.37 , 1.73]), $t(2999) = 17.455$, $p < .001$, vs. good (0.98; 95% CI [0.81, 1.16]), $t(2999) = 11.398$, $p < .001$, ("inflation increase"). Additionally, the effect size for supersession (3c) was smaller when the focal was bad (0.92 ; 95% CI [0.74, 1.10]), $t(2999) = 10.226$, $p < .001$, vs. good (1.49; 95% CI [1.32, 1.66]), $t(2999) = 17.468$, $p < .001$, ("supersession decrease").

3.3 Discussion

We found that the "abnormal increase"/"supersession decrease" pattern identified in Experiment 1 also extends to norms of proper function. Just as for moral norms, the difference in causal ratings of the focal part when the focal part works vs. does not work is greater when the alternate part does not work ("inflation increase"). Conversely, the difference in causal ratings when the alternate part works vs. does

not work is lesser when the focal part does not work ("supersession decrease").

4 General discussion

Two experiments revealed a novel interaction effect of norm violations on causal judgment. First, the experiments replicated two basic phenomena: a focal event is rated as more causal when it is bad ("inflation") and a focal event is rated less causal when the alternative event is bad ("supersession"). Critically, the experiments showed that (1) the difference in causal ratings of the focal event when it is good vs. bad increases when the alternative event is bad ("inflation increase") and (2) the difference in causal ratings of the focal event when the alternative event is bad vs. good decreases when the focal event is bad ("supersession decrease").

Experiment 1 yielded this novel interaction effect in the context of moral norm violations (e.g., stealing a book from the library). Experiment 2 showed that the effect generalized to violations of norms of proper functioning (e.g., a part of a machine working incorrectly).

This interaction pattern is predicted by the necessity/sufficiency model (Icard et al., 2017). The success of this prediction is especially striking, in that the necessity/sufficiency model was not created with this interaction in mind. Rather, the model was originally created to explain inflation and supersession, and it was only noticed later that this model predicts an interaction in cases of this type.

However, regardless of whether the necessity/sufficiency model is in fact correct, the present studies provide strong evidence that this interaction exists. Therefore, there needs to be some sort of explanation for this interaction, whether it comes in the form of a modified version of the necessity/sufficiency model, or from a completely different computational model, or from some other type of explanation that is more qualitative in nature.

4.1 Implications for other theories

Existing research has led to the development of a broad variety of different theories that aim to explain the impact of prescriptive norms on causal judgments (Alicke et al., 2011; Morris et al., 2021; Quillien, 2020; Sytsma et al., 2019; Sytsma, 2021; Samland and Waldmann, 2014, 2016). There is a wide-ranging debate as to which of these theories is correct, and work on this topic has drawn on numerous different sources of evidence (Henne et al., 2021a; O'Neill et al., 2021; Samland and Waldmann, 2016; Sytsma, 2021).

We cannot hope to discuss all of this evidence here (for a review, see Willemsen and Kirfel, 2019, but we do want to explore the question as to what the specific interaction effect observed in our studies might show about each theory. This effect is predicted by the necessity/sufficiency model. Can any of the other theories explain

it? Or, failing that, can they at least introduce auxiliary assumptions that allow them to accommodate it?

The simplest and most straightforward theory would be one that says that causation judgments are "zero-sum," meaning that there is only a fixed amount of causation to go around in total, so that the more people see one factor as a cause, the less they will see any other factor as a cause. The interaction pattern observed here is not compatible with this simple theory. To see why, consider a case in which Suzy and Billy jointly bring about an outcome. The degree to which Suzy's causation increases when she does something bad will be higher when Billy does something bad then when he does something good ("inflation increase"). Thus, any theory according to which there is only a fixed amount of causation to go around would have to say that the degree to which Billy's causation decreases when Suzy does something bad will be higher when Billy does something bad than when he does something good. But the actual empirical results show precisely the opposite pattern. They show that the degree to which Billy's causation decreases when Suzy does something bad is *lower* when Billy does something bad than when he does something good ("supersession decrease"). In short, there seems to be no real way of making sense of this interaction pattern on the assumption that causal judgments are zero-sum.

Of the alternative explanations developed within the existing literature, the two that are most closely comparable to the necessity/sufficiency model are the SAMPLE model (Morris et al., 2021) and the counterfactual effect size model (Quillien, 2020). Much like the necessity/sufficiency model, these two other models generate precise predictions about what causal ratings should be for each possible level of normality of the focal event and each possible level of normality of the alternate event. Both of these other models were also explicitly intended to explain both inflation and supersession. However, it should be noted that the papers introducing these other models both focus almost entirely on explaining the impact of *statistical* norms and only very briefly discuss the extension to prescriptive norms.

To facilitate comparison with the necessity/sufficiency model, we computed the predicted causal ratings for each of these other models in the specific type of causal structure used in the present studies. Although these other models yield different predictions for other types of causal structures, in this particular structure, the predictions of the counterfactual effect size model will always be the square root of the SAMPLE model predictions. Figure 4 shows the model predictions for the SAMPLE model; Figure 5 shows the model predictions for the counterfactual effect size model (CESM). Inspection of the figures shows that both models predict inflation (higher causal ratings at points that are lower on the y-axis) and supersession (lower causal rating at points that are to the right on the x-axis).

However, unlike the necessity/sufficiency model, the SAMPLE and CESM models do not usually predict an interaction effect observed in the present studies (hereinafter "positive" interaction). While the necessity/sufficiency model always predicts a positive interaction, the SAMPLE and CESM models only predict a positive interaction

under certain circumstances. In the Appendix, we discuss the exact circumstances which give rise to instances of a positive interaction under the SAMPLE and CESM models. For example, the SAMPLE model might predict a positive interaction effect, but only under the assumption that the probability of the "focal good" event is lower than the probability of the "alternate good" event (or that the probability of the "focal bad" event is lower than the "alternate bad" event). This would require, for instance, assuming that the probability of Sam taking a pen when allowed is lower than the probability of Brook taking a pen when allowed. Future research could potentially explore whether there really is some systematic asymmetry of this type in vignette studies.

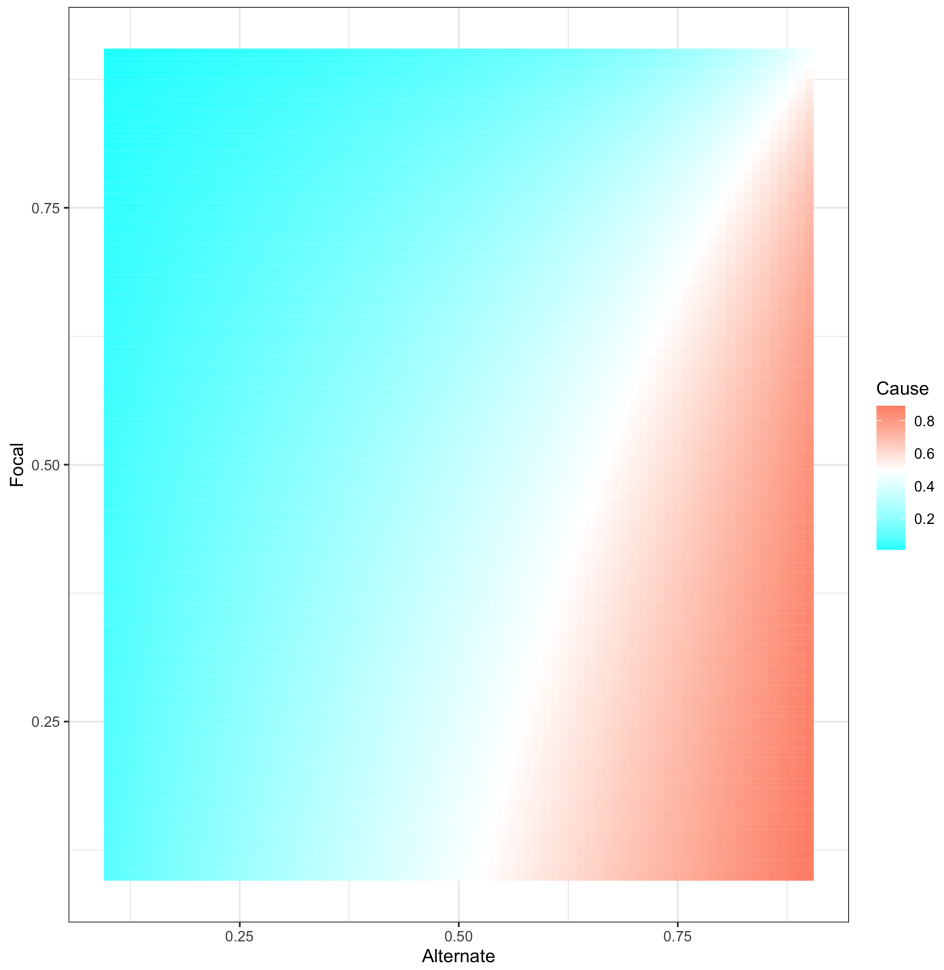

Figure 4: Model predictions for the SAMPLE model (Morris et al., 2021), in the case of an unshielded collider structure in which there are two variables that are individually necessary and jointly sufficient. Color shows predicted causal judgment. In this particular case, the SAMPLE model says that causal judgments should be predicted by $P(F = 0)P(A = 1)/(P(F = 0)P(A = 1) + P(A = 0))$.

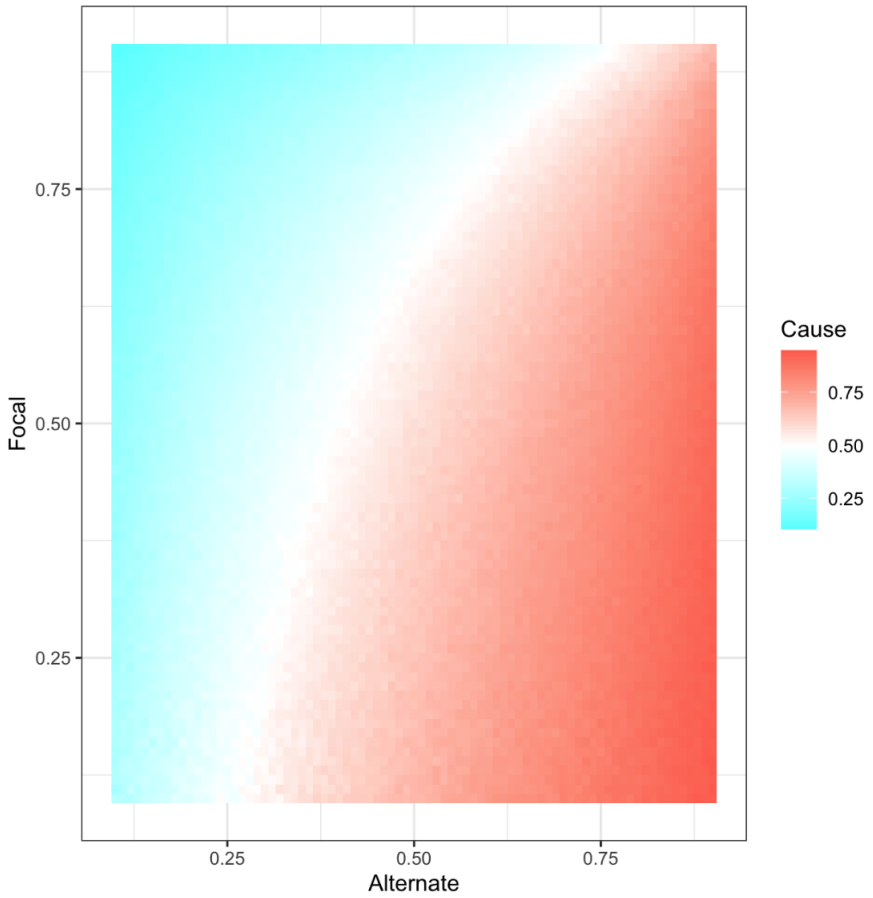

Figure 5: Model predictions for the counterfactual effect size model (Quillien, 2020), in the case of an unshielded collider structure in which there are two variables that are individually necessary and jointly sufficient. Color shows predicted causal judgment. In this particular case, the counterfactual effect size model says that causal judgments can be simulated by (a) sampling the exogenous variables F and A from their prior probabilities, (b) simulating forward to find the value of the effect E and then (c) computing the correlation coefficient between F and E.

Thus, while these models are less likely to predict the observed interaction, it is still possible that the observed interaction in the present studies is occurring because the norm violations meet specific values of "focal good," "focal bad," "alternate good," and "alternate bad." A defender of these models could simply adopt such a position, and make a reasonable inference that these specific assumptions are met in the present studies. Alternatively, a defender of one of these models could also posit that they accurately explain inflation and supersession effects insofar as these effects arise for purely statistical norms and that some further factor is at work in cases involving prescriptive norms like those explored here. We discuss this possibility further in the next section.

Aside from these two models, another proposal in the literature that makes predictions about the role of norms in causal judgment is the account of Halpern and Hitchcock (2015). In its present formulation, this account aims to explain differences in causal judgments for different events within a single vignette, but it does not aim to explain differences between causal judgments across different vignettes. However, if we supplement the account with some additional structure, we get an account that does make predictions about differences across vignettes. A key question, then, will be whether this extended version of the Halpern and Hitchcock account can explain the interaction observed in the present studies.

More specifically, the Halpern and Hitchcock account assumes a "normality ordering" over possible states of each individual causal setup, but it is straightforward to imagine an extension of the account that includes a more quantitative assignment of "degrees of normality" to different possible states, so that the facts about normality all live on a single numerical scale. Halpern and Hitchcock propose that causal strength of a variable will be related to the normality of the counterfactual possibility witnessing that variable's causal status. In the present setting, the relevant counterfactual possibility is always the state in which the second event A still occurred, but the focal event F did not occur. Suppose we label the degrees of normality for this state across the four relevant causal setups as follows:

X: Focal F is good, alternative A is bad.

Y: Both F and A are bad.

Z: Both F and A are good.

W: F is bad, A is good.

As we show in the Appendix, it turns out that the conditions on these numbers that would accommodate our two effects are exactly the same: in both cases we would require that

$$Y + Z - X > W$$

Whether this stipulation can be justified on independent grounds we leave as an interesting open question.

In addition to formal models, a number of qualitative accounts have been proposed to explain the influence of norm violations on causal judgment. First, the "blame validation" hypothesis posits that the effect of moral norm violations on causal judgment stems from a desire to blame bad actors (or praise good actors; Alicke et al. 2011). On this account, an initial negative evaluation of a character and desire to blame that character will distort an existing causal judgment. Second, the "responsibility" account, holds that the dominant use of the word "cause" includes normative considerations such as responsibility and that participants' ratings reflect this dominant meaning rather than a strictly scientific and/or philosophical meaning (Sytsma et al., 2019; Sytsma, 2021). Third, the "pragmatics" account argues that the results found

in the literature may stem from task demands: specifically, participants could be inferring that the experimenter is not really asking about causation, but is instead asking about a different concept, such as the agent's accountability (Samland and Waldmann, 2014, 2016).

Though these alternative qualitative accounts are quite different from each other, they share a common theme in that they suggest that the patterns observed in people's causation judgment in fact reflect patterns in their attributions of blame, responsibility, or accountability. A key question now is whether this broad approach has the potential to explain the specific pattern observed in the present studies. Does this specific interaction pattern also arise for attributions of blame, responsibility, or accountability?

To see whether the pattern we observed for causation judgments is mirrored in the pattern of blame attributions, we would have to look at the difference between the blame an agent receives when she did something bad and the blame an agent receives when she did not do anything bad. Then that difference would have to be *greater* in the case where another agent did something bad than in the case where the other agent did not do anything bad. Further research could explore the question as to whether this pattern actually arises.

4.2 Generalizing to statistical norms

The present studies focused exclusively on the impact of prescriptive norms, such as moral norms (Experiment 1) and norms of proper functioning (Experiment 2). However, long before researchers began exploring the impact of prescriptive norms, it was well known that causal judgments could be impacted by purely statistical norms, e.g., that people's judgment about whether a particular factor caused an outcome could be impacted by the degree to which they saw that factor as statistically frequent or infrequent (Kelley, 1967; Hilton and Slugoski, 1986). Within more contemporary research on this topic, one key question has been whether the impact of statistical norms on causal judgments shows the same patterns found for the impact of prescriptive norms (Icard et al., 2017; Samland and Waldmann, 2014; Sytsma et al., 2012). The interaction observed in the present studies could potentially shed further light on this issue.

In particular, the necessity/sufficiency model predicts that precisely the same interaction should arise for statistical norms. A question now arises as to whether that prediction will turn out to be correct. There are two main classes of methodologies to approach this prediction: first, vignette methods (akin to the present studies); second, more quantitative methods (explained below).

4.2.1 Vignette methods

In studies using vignette methods, probabilistic information is presented to participants in a vignette using the sorts of ordinary intuitive phrases people use to talk

about frequencies in everyday life. (For example, participants might be given a story about a person who takes a pen from the office and then told that such people "typically do take pens"; Roxborough and Cumby 2009.) Research using this method does find abnormal inflation for statistical norms, but the effect is much smaller than for moral norms. In one recent vignette study, for example, the unstandardized effect size (b) for the inflation effect for statistical norms was less than half the size of the inflation effect for moral norms (Icard et al., 2017).

Because effect sizes are so low, this vignette-based method might prove impracticable to test for an interaction in the domain of statistical norms. To get a sense for the required sample size, we conducted a post-hoc power analysis of the interaction effect observed for moral norms in our Experiment 1 (by resampling from the data obtained in that study). The results show that approximately 1500 participants would be needed for 80% power and 2400 for 95% power within the realm of moral norms. If the effect size for statistical norms turns out to be less than half the effect size for moral norms, the sample size required to detect the interaction for statistical norms would be considerably larger.

One possibility for future researchers would be to pursue a vignette approach in a within-subjects study. Theoretically, this would allow for more observations without having to recruit an extraordinarily large number of participants, assuming that the effect size remains constant. However, some prior related work in experimental philosophy has suggested that the effects of norm violations on people's judgments is sensitive to the number of vignettes that participants have to respond to. For example, in a study of intentional action judgments, Cushman and Mele (2008) presented participants with a series of vignettes (in counterbalanced order) and found that judgments for the last vignette were much less sensitive to norm violations than were judgments for the first vignette.

In light of this, we face a genuine question as to whether it is even feasible to test for an interaction with statistical norms using this same method. One view would be that it is not feasible and we need to turn to a different method, while another would be that it is feasible and simply requires either a within-subject design or an extraordinarily large sample size.

4.2.2 Quantitative methods

Another class of methods is more directly quantitative in nature (e.g., Morris et al. 2019; cf. Gerstenberg and Icard 2020; Kirfel et al. 2021). In studies using these methods, participants are presented with information that can be used to estimate the probabilities of particular events (either by giving participants explicit information about the probabilities or by giving them a sample from the relevant distribution). Studies using these methods have found both inflation and supersession for statistical norms. (e.g., Morris et al. 2019; cf. Gerstenberg and Icard 2020; Kirfel et al. 2021)

In one such study, Morris and colleagues (2019) gave participants information

about the probabilities of different events by displaying urns that contained balls of different colors. On each trial, participants were given information about the probability of the focal event and the probability of the alternate event. Each participant completed five trials, with five different pairs of probabilities.

Because Morris et al. independently manipulated the probability of the focal and the probability of the alternate, it is possible to test whether our interaction effect occurs in their data. We therefore conducted an additional analysis on their existing dataset. To determine whether there was an interaction effect, we ran a mixed effects model predicting the causal rating, treating probability of focal (centered) and probability of alternate (centered) as fixed effects and participant as a random factor with respect to intercept. This model did not find a significant interaction between focal and alternate, $b = .003$; 95% CI: [-0.010, 0.004], $\chi^2(1, 4953) = 0.578$, $p = .55$. Also, as indicated in the Appendix, we ran analyses to determine which "quadruple" of values would yield the largest interaction effect according to the necessity/sufficiency model (focal bad and good at 0.1 and 0.9 respectively and alternate bad and alternate good at 0.1 and 0.9). Subsetting the data to these probabilities only and bootstrapping therein did not reveal an interaction effect, either.

That there is no interaction effect in the Morris et al. data raises two possibilities. One possibility would be that there is no interaction in these data because the interaction effect exists for prescriptive norms but not statistical norms. A second possibility would be that there is no interaction in these data because of methodological differences between the present studies and the Morris task.Though existing work has not directly tested this hypothesis directly with regard to causal judgments, some studies of causal inference show that when participants are randomly assigned to receive the information in vignette or quantitative format, the impact of prescriptive norms seems to arise only in the vignette format (Danks et al., 2014; Samland and Waldmann, 2014). Thus, it is an open question whether the effect observed in the present studies will also occur for statistical norms.

Future work could explore these issues at a number of different levels. At a deeper level, there is the question as to why different experimental methodologies yield such different patterns of results. Here, defenders of different theories will presumably give very different answers. Defenders of the pragmatic theory might say that participants who get the information in the form of a vignette feel pragmatic pressure to take prescriptive information into account (Samland and Waldmann, 2014), whereas defenders of the necessity/sufficiency theory might say that participants who get the information in a more quantitative format will tend to think of the task almost as a kind of mathematical puzzle and therefore stop using their more intuitive way of thinking about causation. This is a very fundamental question, but perhaps not one that will be resolved in the near future.

On another level, however, we face a more straightforward question regarding the interaction observed here. The present studies show that this interaction arises for prescriptive norms. We can now ask: If participants receive information in a format

that does yield the interaction observed here for prescriptive norms, will they also show an interaction for statistical norms? This is a difficult question—with potential pitfalls both for vignette methods and for more quantitative methods—but at the very least, it does seem like the sort of question that might be conclusively resolved in the short term through future research.

4.3 Conclusion

The present studies provide evidence for a novel interaction effect concerning the influence of prescriptive norm violations on causal judgments. Previous work found two main effects: that norm violations of the focal and alternate events both influence judgments of the focal event (i.e., "inflation" and "supersession"). We build on this literature by showing that these two main effects interact in a specific way (i.e., "inflation increase" and "supersession decrease").

The necessity/sufficiency model (Icard et al., 2017) specifically predicts our novel interaction, and the results therefore provide evidence that this model might be at least broadly on the right track. More generally, the existence of the interaction effect creates the opportunity for future model building and refinement. In the past few years, a number of theories, both quantitative and qualitative, have focused on understanding the effect of norm violations on causal judgment. Beyond the necessity/sufficiency model explanation, we are curious about other alternative explanations for the interaction effect, and look forward to how this interaction effect might contribute to future theory-building and research.

Acknowledgements

We would like to thank the reviewers for their insightful analysis and contributions, which have significantly changed this paper.

References

Alicke, M., Rose, D., and Bloom, D. (2011). Causation, norm violation, and culpable control. *Journal of Philosophy*, 108(12):670–696.

Alicke, M. D. (1992). Culpable causation. *Journal of personality and social psychology*, 63(3):368.

Bear, A. and Knobe, J. (2017). Normality: Part descriptive, part prescriptive. *cognition*, 167:25–37.

Blanchard, T. and Schaffer, J. (2017). Cause without default. *Making a difference*, pages 175–214.

Cushman, F., Knobe, J., and Sinnott-Armstrong, W. (2008). Moral appraisals affect doing/allowing judgments. *Cognition*, 108(1):281–289.

Danks, D., Rose, D., and Machery, E. (2014). Demoralizing causation. *Philosophical Studies*, 171(2):251–277.

Driver, J. (2008). Attributions of causation and moral responsibility. *Moral Psychology*, 2:423–440.

Gerstenberg, T. and Icard, T. (2020). Expectations affect physical causation judgments. *Journal of Experimental Psychology: General*, 149(3):599–607.

Halpern, J. Y. and Hitchcock, C. (2015). Graded causation and defaults. *British Journal for the Philosophy of Science*, 66(2):413–457.

Henne, P., Kulesza, A., Perez, K., and Houcek, A. (2021a). Counterfactual thinking and recency effects in causal judgment. *Cognition*, 212:104708.

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

Henne, P., Pinillos, Á., and De Brigard, F. (2016). Cause by omission and norm: Not watering plants. *Australasian Journal of Philosophy*, pages 1–14.

Hilton, D. J. and Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological review*, 93(1):75.

Hitchcock, C. and Knobe, J. (2009). Cause and norm. *Journal of Philosophy*, 106(11):587–612.

Icard, T. (2016). Subjective probability as sampling propensity. *Review of Philosophy and Psychology*, 7(4):863–903.

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

Kelley, H. (1967). Attribution theory in social psychology.

Kirfel, L., Icard, T., and Gerstenberg, T. (2021). Inference from explanation. *Journal of Experimental Psychology: General*. https://doi.org/10.31234/osf.io/x5mqc.

Knobe, J. and Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. *Moral Psych.*, 2:441–448.

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

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

Lieder, F., Griffiths, T. L., and Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological review*, 125(1):1.

Livengood, J., Sytsma, J., and Rose, D. (2017). Following the FAD: Folk attributions and theories of actual causation. *Review of Philosophy and Psychology*, 8(2):273–294.

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Morris, A., Phillips, J., Gerstenberg, T., and Cushman, F. (2019). Quantitative causal selection patterns in token causation. *PLoS ONE*, 14(8).

Morris, A., Phillips, J. S., Icard, T., Knobe, J., Gerstenberg, T., and Cushman, F. (2021). Causal judgments approximate the effectiveness of future interventions.

O'Neill, K., Henne, P., Pearson, J., and De Brigard, F. (2021). Measuring and modeling confidence in human causal judgment.

Phillips, J., Luguri, J. B., and Knobe, J. (2015). Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition*, 145:30–42.

Phillips, J., Morris, A., and Cushman, F. (2019). How we know what not to think. *Trends in cognitive sciences*, 23(12):1026–1040.

Quillien, T. (2020). When do we think that X caused Y? *Cognition*, 205.

Roxborough, C. and Cumby, J. (2009). Folk psychology concepts: Causation 1. *Philosophical Psychology*, 22(2):205–213.

Samland, J., Josephs, M., Waldmann, M. R., and Rakoczy, H. (2016). The role of prescriptive norms and knowledge in children's and adult's causal selection. *Journal of Experimental Psychology: General*, 145(2):125–130.

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

Samland, J. and Waldmann, M. R. (2014). Do social norms influence causal inferences? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36.

Sanborn, A. N. and Chater, N. (2016). Bayesian brains without probabilities. *Trends in cognitive sciences*, 20(12):883–893.

Sytsma, J. (2021). The responsibility account. *Advances in Experimental Philosophy of Causation*.

Sytsma, J., Bluhm, R., Willemsen, P., and Reuter, K. (2019). Causal attributions and corpus analysis. *Methodological advances in experimental philosophy*, pages 209–238.

Sytsma, J., Livengood, J., and Rose, D. (2012). Two types of typicality: Rethinking the role of statistical typicality in ordinary causal attributions. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(4):814–820.

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

Appendix

Main Predictions and Bounds

In this brief appendix we explain in more detail some of the technical claims mentioned in the main text. Chief among these is the derivation of the two primary predictions from the necessity/sufficiency account, that is, inflation increase and supersession decrease. We also explain how the interaction effect explored in the present paper relates to the two main effects explored in previous work, namely, inflation and supersession.

Recall that in the conjunctive cases we study here the necessity/sufficiency account assigns a causal strength of

$$P(F=1)P(A=1)+1-P(F=1). \tag{1}$$

Inflation, supersession, inflation increase, and supersession decrease all involve contrasts between cases when a term—either $P(F=1)$ or $P(A=1)$—is high and cases when it is low. Suppose we are dealing with two scenarios in which $P(F=1)$ is high and low, and let us denote these values respectively by p_F^+ and p_F^-. We adopt analogous abbreviations for two possible values of $P(A=1)$, namely p_A^+ and p_A^-. Then, for instance, given a fixed value p_A for $P(A=1)$, the degree of inflation (the difference between the two instances of (1) when $P(F=1)$ is low and when it is high) will be:

$$(p_F^- p_A + 1 - p_F^-) - (p_F^+ p_A + 1 - p_F^+) \;=\; (p_F^- - p_F^+)(p_A - 1). \tag{2}$$

Similarly, fixing a particular value p_F for $P(F=1)$, the degree of supersession is predicted to be:

$$(p_F p_A^+ + 1 - p_F) - (p_F p_A^- + 1 - p_F) \;=\; (p_A^+ - p_A^-)(p_F - 1). \tag{3}$$

Inflation increase concerns the difference in (2) when $P(A = 1)$ is low and when it is high, while supersession decrease concerns the difference in (3) when $P(F = 1)$ is high compared to when it is low. Thus, the degree of inflation increase is predicted to be:

$$\begin{aligned}(p_F^- - p_F^+)(p_A^- - 1) - (p_F^- - p_F^+)(p_A^+ - 1) &= (p_F^- - p_F^+)(p_A^- - p_A^+) \\ &= (p_F^+ - p_F^-)(p_A^+ - p_A^-),\end{aligned}$$

where the second equality follows from the fact that these numbers all fall in the unit interval between 0 and 1. The degree of supersession decrease is also predicted to be:

$$(p_A^+ - p_A^-)(p_F^+ - 1) - (p_A^+ - p_A^-)(p_F^- - 1) = (p_F^+ - p_F^-)(p_A^+ - p_A^-)$$

In fact, independent of the necessity/sufficiency account, the degree of inflation increase is always guaranteed to be the same as the degree of supersession decrease.

Abbreviating $\Delta_F = p_F^+ - p_F^-$ and $\Delta_A = p_A^+ - p_A^-$, we have just seen that, according to the necessity/sufficiency account, both inflation increase and supersession decrease correspond to a specific quantity δ:

$$\delta = \Delta_F \Delta_A.$$

That is, δ is only as great as the smallest difference in probability values, and we only reach that quantity when the other difference is close to maximal, that is, when the larger and smaller values are close to 1 and 0, respectively. This of course puts a sharp upper bound on any effect we would expect to find. At the same time it puts some lower bound on δ. In particular, as long as Δ_F and Δ_A are both positive (in other words, as long as the experimental manipulation is successful), the interaction δ is guaranteed to be positive.

To explore further bounds on Δ, suppose we consider the *inflation main effect*, which can understood as averaging the inflation effect (2) over the two (high and low) values of $P(A = 1)$. (In our experiments to determine the main effect, half of the participants were in the condition with $P(A = 1)$ high, and half with $P(A = 1)$ low.) According to the necessity/sufficiency account, this would correspond to the following value, abbreviating $\mu_A = (p_A^+ + p_A^-)/2$:

$$\begin{aligned}\iota &= \tfrac{1}{2}((p_F^- p_A^+ - p_F^- + 1) - (p_F^+ p_A^+ - p_F^+ + 1)) + \\ &\quad \tfrac{1}{2}((p_F^- p_A^- - p_F^- + 1) - (p_F^+ p_A^- - p_F^+ + 1)) \\ &= \tfrac{1}{2}(p_F^- p_A^+ - p_F^- - p_F^+ p_A^+ + p_F^+) + \tfrac{1}{2}(p_F^- p_A^- - p_F^- - p_F^+ p_A^- + p_F^+) \\ &= (p_F^- p_A^+ - p_F^- - p_F^+ p_A^+ + p_F^+ + p_F^- p_A^- - p_F^- - p_F^+ p_A^- + p_F^+)/2 \\ &= (p_F^-(p_A^+ + p_A^- - 2) + p_F^+(2 - p_A^+ - p_A^-))/2 \\ &= ((p_F^+ - p_F^-)(2 - p_A^+ - p_A^-))/2 \\ &= \Delta_F(\frac{2 - (p_A^+ + p_A^-)}{2}) \\ &= \Delta_F(1 - \mu_A)\end{aligned}$$

Likewise for the *supersession main effect* (or "average supersession effect") when we equally weight the two effect sizes (3) by the high and low values for $P(F = 1)$. Abbreviating $\mu_F = (p_F^+ + p_F^-)/2$ we then have:

$$\begin{aligned}
\sigma &= \tfrac{1}{2}((p_F^+ p_A^+ - p_F^+ + 1) - (p_F^+ p_A^- - p_F^+ + 1)) + \\
&\quad \tfrac{1}{2}((p_F^- p_A^+ - p_F^- + 1) - (p_F^- p_A^- - p_F^- + 1)) \\
&= (p_F^+ p_A^+ - p_F^+ - p_F^+ p_A^- + p_F^+ + p_F^- p_A^+ - p_F^- - p_F^- p_A^- + p_F^-)/2 \\
&= (p_F^+(p_A^+ - p_A^-) + p_F^-(p_A^+ - p_A^-))/2 \\
&= (\Delta_A(p_F^+ + p_F^-))/2 \\
&= \Delta_A \mu_F
\end{aligned}$$

The inflation main effect ι and the supersession main effect σ together give us a slightly better lower bound on δ. While we know $\delta > 0$, we also see that

$$\begin{aligned}
\delta &= \Delta_F \Delta_A \\
&\geq \Delta_F \Delta_A (1 - \mu_A)\mu_F \\
&= \iota\sigma.
\end{aligned}$$

That is, the necessity/sufficiency model predicts that the interaction effect will always be greater than or equal to the inflation main effect multiplied by the supersession main effect. For example, if the inflation main effect is .3 and the supersession main effect is .2, then the necessity/sufficiency model predicts that the interaction effect must be greater than or equal to .06.

Significantly, both main effects, ι and σ, could be quite high, while δ is still rather low. For instance, when $p_F^+ = .95$, $p_F^- = .65$, $p_A^+ = .35$, and $p_A^- = .05$, then both main effects would be at least 0.24, while δ is merely 0.09. (In this case, the lower bound $\iota\sigma$ noted above would tell us that it must be at least 0.05796.) Needless to say, detecting such a small effect, even if present, could require a very large sample size..

Predictions Based on Degrees of Normality

As in the main text, let us label four relevant scenarios as follows:

X: Focal F is good, alternative A is bad.

Y: Both F and A are bad.

Z: Both F and A are good.

W: F is bad, A is good.

That is, for example, X is a numerical value representing a "degree of normality" for the situation in which F is good, A is bad, but both F and A happen.

A causal variable, such as F, will receive strength proportional to a counterfactual possibility that witnesses F's causal status. In the simple cases we are considering here, where F and A were conjunctively necessary to bring about the outcome, this is the possibility in which F did *not* occur (as it in fact did). In this setting the inflation effect would be predicted to the extent that $W > Z$ and $Y > X$. The best witness for the case in which F was bad should be more normal than the best witness for the case in which F was good (independent of A's status). In other words, both $W - Z$ and $Y - X$ should be positive. To predict the interaction we would need to compare these two differences: the effect should be larger when the alternative A is bad. That is, we need $Y - X > W - Z$, or, as we put it in the main text:

$$Y + Z - X > W \tag{4}$$

Next, consider supersession. This pattern is predicted to the extent that $W > Y$ and $Z > X$. But supersession descrease would be involve comparisons these two differences. We would need $Z - X > W - Y$. Once again, this is precisely the same pattern (4) that we would need for inflation increase.

Summarizing, an account in the spirit of Halpern & Hitchcock (2015) based on normality degrees for counterfactual witnesses to causal status would be compatible with the effects we have shown in this paper. Specifically, we need the arithmetical relationship in (4) to hold among the four normality values at issue.

Example prediction of Necessity/Sufficiency

As mentioned in the main text, the necessity/sufficiency model predicts an interaction effect characterized by inflation increase/supersession decrease. See Figure 4 for an illustration of what the necessity/sufficiency model predicts for the case in which the probabilities of focal good and alternate good are both .6 and the probabilities of focal bad and alternate bad are both .4.

Comparing the predictions of SAMPLE, CESM and Necessity/Sufficiency

In this section, we explore the conditions under which the SAMPLE model and the counterfactual effect size model (CESM) would predict the interaction pattern observed in the present studies. As discussed in the previous section, the necessity/sufficiency model always predicts such an interaction.

To examine this issue computationally, we began by constructing a set of cases for which we could generate predictions. Each case consisted of a quadruple of probabilities (focal good, focal bad, alternate good, alternate bad). Probabilities were assigned in increments of .1, and we only considered cases such that the probability for focal good was higher than the probability for focal bad, and the probability for alternate

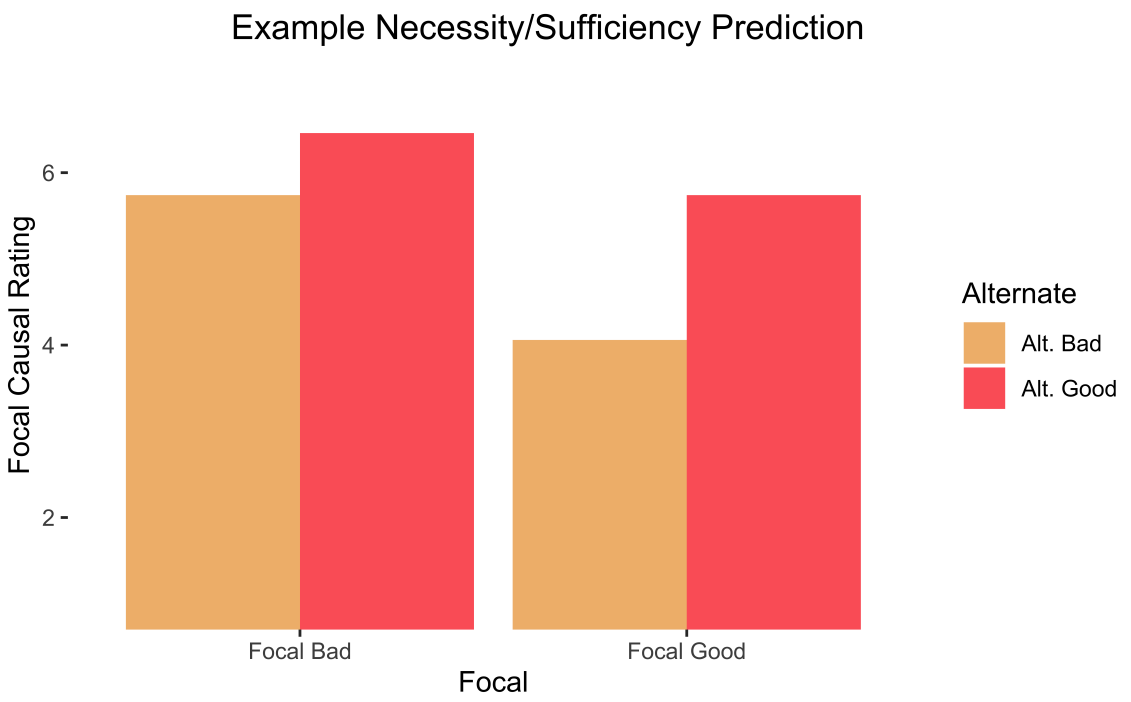

Figure 6: Example prediction arising from the necessity/sufficiency model, for the case in which the probabilities of focal good and alternate good are both .7 and the probabilities of focal bad and alternate bad are both .3. For purposes of interpretation, model predictions were converted from a 0-1 scale to a 1-7 scale and displayed as such.

good was higher than the probability for alternate bad. In total, 1944 quadruples of probabilities fulfilled these constraints.

For each model, we computed the predicted effect size of the interaction for each quadruple (with higher positive numbers corresponding to a larger effect size in the direction observed in the present studies). Figure 1 shows the predicted interaction effect according to SAMPLE and CESM, both contrasted with the predictions of the necessity/sufficiency model. As the figure shows, necessity/sufficiency predicts a positive interaction for all quadruples (100%). By contrast, the CES and SAMPLE models predict a positive interaction less often (47.11% and 29.57%, respectively).

We then sought to understand the conditions under which SAMPLE and CESM would predict a positive interaction effect. In particular, we looked at the difference between the probability of focal good and the value of alternate good ("good distance") and the difference between the probability of focal bad and alternate bad ("bad distance").

Figure 2 shows how the predictions of the models vary with respect to differences between the corresponding "focal" and "alternate" conditions. We consider the distances for the "good" variables and for the "bad" variables. As can be seen in Figure 2, for both the SAMPLE and CESM, good distance is strongly determinative

of whether there is a positive interaction effect. Bad distance also appears to be influential.

Finally, we looked at cases that were "symmetric," in the sense that the probability of focal good was equal to the probability of alternate good and the probability of focal bad was equal to the probability of alternate bad. As Figure 3 shows, in these symmetric cases, SAMPLE and CESM predict either that there should be no positive interaction or that interaction should be of a negligible size. By contrast, necessity/sufficiency always predicts positive interactions, and sometimes interactions of a substantially larger size.

Analysis of Morris et al. (2019) task data

Morris et al. (2019) investigated how people use information about probabilities to make causal judgments. On each trial, participants were given information about the probability of the focal event and the probability of the alternate event. Participants made a rating about the causal strength of the focal event. Each participant completed five trials, with five different pairs of probabilities.

Because Morris et al. independently manipulated the probability of the focal and the probability of the alternate, it is possible to test whether our interaction effect occurs in their data. We therefore conducted additional analyses. To test for a possible interaction effect, we first ran a mixed effects model predicting the causal rating, treating probability of focal (centered) and probability of alternate (centered) as fixed effects and participant as a random factor with respect to intercept. This model did not find a significant interaction between focal and alternate, $b = .003$; 95% CI: [-0.010, 0.004], $\chi^2(1, 4953) = 0.578$, $p = .55$.

Additionally, in our prior computational analysis of predicted interaction effect sizes, we can now select a "quadruple" that would yield the largest interaction effect according to the necessity/sufficiency model: focal bad and good at 0.1 and 0.9 respectively and alternate bad and alternate good at 0.1 and 0.9. Subsetting the Morris data to include only those trials (N = 212) and resampling with replacement to create a dataset of the original size (N = 4964), there was again no significant interaction effect using the same mixed effects model, $b = 0.599$; 95% CI: [-0.792, 1.991], $\chi^2(1, 4953) = 0.718$, $p = .396$. For a discussion of the implications of these analyses, see the General Discussion.

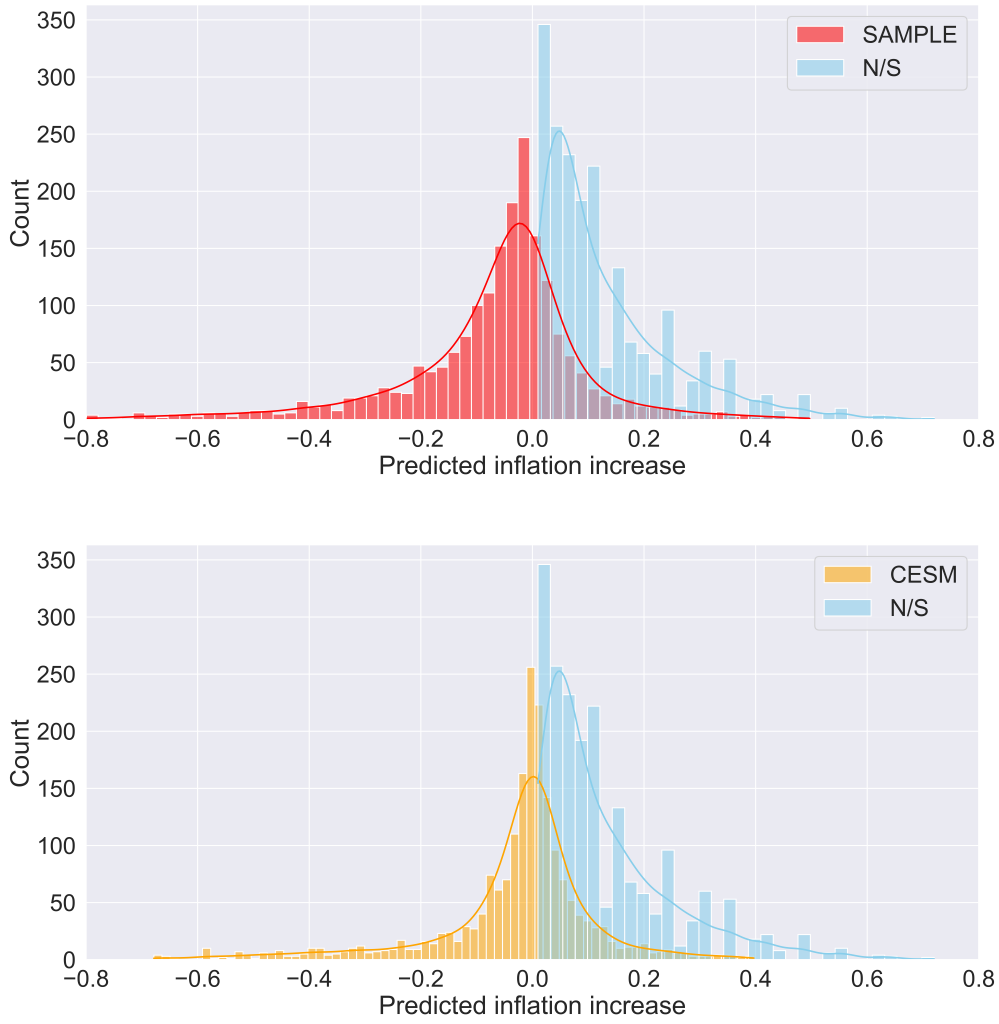

Figure 7: Counts of model-predicted interaction effects by magnitude and direction. For both figures, the X-axis represents the size and direction of predicted inflation increase, and the Y-axis represents counts (of 1944 quadruples.) Above: predictions of the SAMPLE model (red) contrasted with necessity/sufficiency (blue). Below: predictions of CESM (orange) contrasted with necessity/sufficiency (blue).

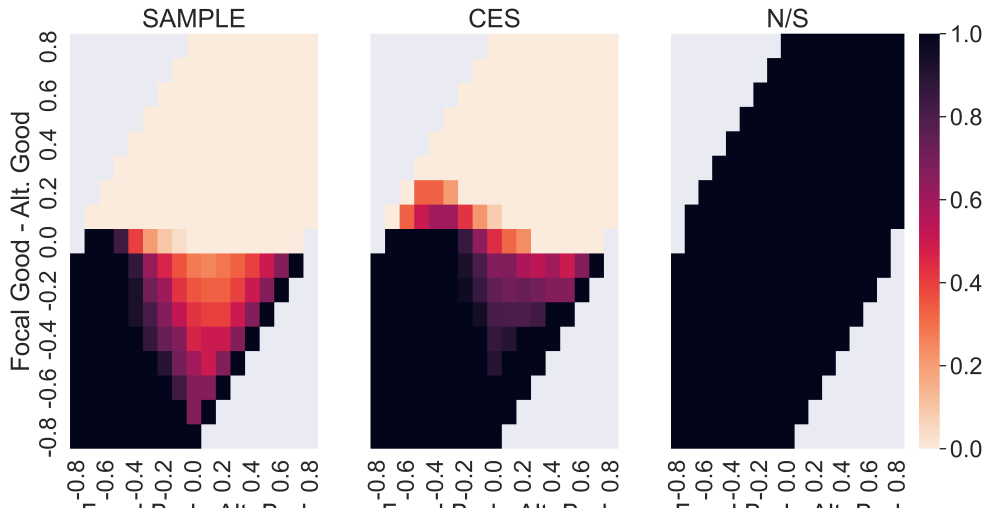

Figure 8: Model predictions by good distance and bad distance. The subplots depict whether model predictions for SAMPLE, CESM, and necessity/sufficiency (left, center, and right) are always positive (black) or not. X and Y axes depict bad distance and good distance, respectively.

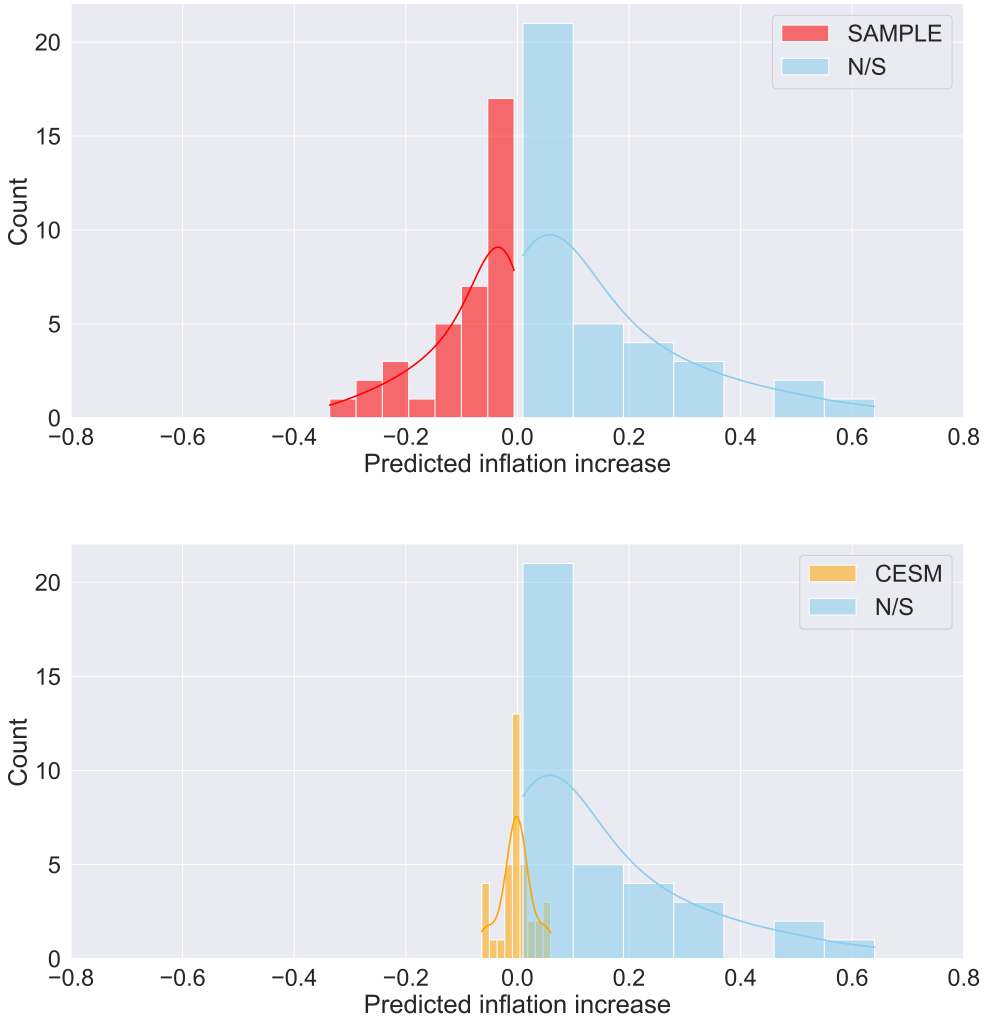

Figure 9: For "symmetric" quadruples only, counts of model-predicted interaction effects by magnitude and direction. As in Figure 1, for both figures, the X-axis represents the size and direction of predicted inflation increase, and the Y-axis represents counts. Above: predictions of the SAMPLE model (red) contrasted with necessity/sufficiency (blue), assuming symmetry. Below: predictions of CESM (orange) contrasted with necessity/sufficiency (blue), also assuming symmetry.