Normal Causation

Songzhi Wu, Catherine Holland and others and
Jonathan Phillips

2024

Intro:

In the critically acclaimed sitcom *Arrested Development*, Lucille Bluth's thirty-two year-old son, Buster, went swimming in the ocean in a brief moment of rebellion against his mother, after he discovered that she had been hiding the true identity of his father from him. In the ocean, Buster encountered a loose seal that had been freed by his magician brother, Gob, after a failed magic show. During this encounter, the seal bit off Buster's left hand. In such a case, one may have many questions, such as: what was the cause of Buster losing his hand? One may attribute Buster's loss of hand to his imprudent behavior of going into the ocean, the fact that his mother never told him who his true father was, or perhaps the improbable chance encounter with a loose seal. Just like this, we perform causal selection in our daily life.

There are two types of norms that are important to causal cognition: descriptive and prescriptive norms. Descriptive norms refer to how people generally behave while prescriptive norms refer to how people believe we should behave. Prior research has found that both types of norms affect how we think of the cause of an outcome. There has been a number of competing explanations for this effect. For descriptive norms, most researchers believe that the effect is best accounted for by the standard processes of causal reasoning. However, for prescriptive norms, opinions differ. Some believe that the effect simply stems from the responsibility judgment, especially when the outcome is adverse, while others have argued for a more general account emphasizing the general role of normality in causal judgments for both descriptive and prescriptive norms.

We bring in new perspectives to the debate. First, we compare two kinds of prescriptive norms: rationality and morality. Second, we investigate the importance of whether the agent is knowledgeable about violating a norm. Third, unlike prior work, we separate the normality of the action and the valence of the outcome, and we ask whether the effect of the former changes depending on the latter. Fourth, we examine whether descriptive normality of the outcome impacts causal judgments. Accordingly, we add four new findings to the field. First, our studies provide evidence in favor of the general normality account as the two prescriptive norms - rational norms and moral norms - exert similar influence in causal judgment. Second, agent's ignorance about norm violation affects both kinds of norms similarly. Third, both effects are moderated by the valence of the outcome as well as the descriptive normality of the outcome. Fourth, both effects discussed above can be explained by the counterfactual judgment, which provides further support to the theory that normality affects causal judgments by affecting counterfactuals.

Descriptive Norms

Descriptive norms, including statistical norms, shape judgments of causal selection; specifically, events which violate descriptive norms (e.g., low probability events) are often selected as the cause of later events that depended on their occurrence. It has long been discussed that causal judgments may be sensitive to norm violations or expectations about what will occur (Gorovitz, 1965; Hart & Honoré, 1985; Hilton & Slugoski, 1986; Kahneman & Miller,1986). More recent studies delve into the mechanisms of statistical norm violation

in cognition and several models have been proposed to account for the effect. For instance, the counterfactual simulation model (CSM) constructed by Gerstenberg and colleagues shows that the difference between the counterfactual considered and what in fact happened factors into people's causal judgment (2014). Later, researchers expanded the model, arguing that when an unexpected action led to a favorable outcome, the actor would be given credit but when an unexpected action led to an unfavorable outcome, the actor would be assigned blame (Stephan et al., 2017). Furthermore, more credit would be given or more blame would be assigned to agents who are believed to be dispositionally good or bad at acting optimally because when predicting future behavior, people make inferences about agents based on the action history of the latter (Gerstenberg et al., 2017). More specifically, researchers have argued that statistical normality exerts influence on people's causal judgment through probabilistic sampling (Hitchcock & Knobe, 2009). And people's understanding of norms, such as the strength of norm violation, also plays an important role in the causal calculation (Icard et al., 2017). A key concept, counterfactual potency, is constructed to measure the "strength and impact of counterfactual" and has performed well in predicting the impact of counterfactual reasoning in causal judgments (Petrocelli et al., 2011). In addition, the effect of assigning more causation to low probability events is present regardless of the valence of the outcome and it also affects causal attribution to other agents, while higher frequency of norm violation are associated with increased causal attribution (Kominsky et al., 2015; Kirfel & Lagnado, 2017).

Prescriptive Norms

Prescriptive norms, including moral norms and rational norms, have been shown to have similar effects as descriptive norms on causal selection: immoral actions, for example, tend to be selected as the causes of later events that depended on their occurrence (Alicke, 1992). Moreover, moral judgments have been shown to exert influence on causal cognition, instead of only the other way around (Knobe & Fraser, 2008; Knobe, 2010). Important models proposed for moral norms include the culpable control model, the counterfactual reasoning in causal selection model and the accountability hypothesis (Alicke, 2000; Samland & Waldmann, 2016). The three models attempt to decipher the causal attribution by, respectively, referring to people's exaggeration of the causal strength of moral norm violation, tendency to consider abnormal counterfactuals over normal ones, and propensity to consider factors that are present in moral reasoning generally (Samland & Waldmann, 2016). However, there is no consensus in sight on which account is the most favorable. And as in the case of descriptive norms, the relevance of moral norms is quite important in the causal reasoning process (Phillips et al., 2015).

Normality

Central to constructing a unified account for causal judgment is the concept of normality. In the current context, normality means alignment with norms, either descriptive or prescriptive (Halpern & Hitchcock, 2015; Icard et al., 2017). Descriptive normality concerns the probability of the event's occurrence while prescriptive normality may concern the righteousness, legality or reasonableness of the action, depending on the relevant prescriptive norm in the given situation. Actions that are not in accordance with either descriptive or prescriptive norms are thus termed "abnormal". In our experiments below, we will investigate the impact of normality in causal cognition across different types of norms by way of affecting the counterfactuals that come to people's minds.

Ongoing Debate

There is an ongoing debate about whether the two effects discussed above should be understood as arising from the normal process by which people make causal judgments. On the one hand, many have argued that they should not, and that the effect of moral norms or statistical norms are not part of the process of causal reasoning in the first place. Rather, the significant element in the causal reasoning is the valence of the situation or ascription of responsibility (Samland & Waldman, 2016; Alicke et al., 2011; Livengood et al., 2017). On the other hand, a number of researchers have argued that the impact of both descriptive and

prescriptive norms should be understood as part of the normal process of causal judgments and counterfactual structure by appealing to the role of counterfactuals in causal cognition. Supporters of the second theory suggest that people's consideration of relevant counterfactuals may be the basis of a unified account of causal judgments (Phillips et al., 2015; Phillips & Knobe, 2018).

Using empirical experiments reported below, we contribute to the existing debate in several ways. First, we consider a new kind of norm violations, rational norm violations, and find that they have a similar impact as moral norm violations on causal selection (see Johnson & Rips, 2015 and Halpern & Hitchcock, 2015 for previous work on what effect violating rational norms has on causal judgments). Second, we replicate the previously demonstrated outcome moderation effects where negative outcomes are more associated with causal attribution than positive outcomes. but also find that they occur for rational norm violations, further prompting the search for a theory that is not specific to moral norm violations. Third, we find that all of the above-mentioned effects are mediated by participants' counterfactual judgments. Fourth, we Fifth, we demonstrate that these outcome effects are not driven by the valence of the outcome but rather by the normality of the outcome by showing that the same pattern is more frequently observed in the cases of abnormal outcomes than normal outcomes. Finally, we test three different accounts of these effects, one that depends only on the morality of the events, one that depends only on the probability of the events, and one that depends on the normality of the events. We find that the normality accounts best captures people's causal judgments, which calls for a unifying model of causal judgments that is grounded in the normality of relevant events.

Experiment 1: Characterizing moral norm violations in causal chains

To begin our investigation in causal reasoning about agents and inanimate objects, we started by attempting to replicate the effect of norm violations previously found in the literature in a new set of scenarios involving causal chains (see *Fig* 1.). We used 24 scenarios involving a distal cause, which brings about some immediate outcome, which then in turn contributes to a subsequent, final outcome. In these scenarios, we varied the type of distal cause (knowledgeable human agent, ignorant human agent, or inanimate object) and the valence of the final outcome (good or bad). Critically, this design allowed us to replicate prior work by asking whether agents who knowingly bring about bad outcomes are judged to be more causal than agents who knowingly bring about good outcomes. In addition, we can compare this effect of outcome for knowledgeable agents to (1) agents who unknowingly bring about good vs. bad outcomes, and to (2) inanimate objects that make the same causal contribution.

Across all of these conditions, we ask participants both to make judgments of whether the distal cause "caused" the final outcome and to make a counterfactual judgments about what would would have prevented the outcome from occurring.

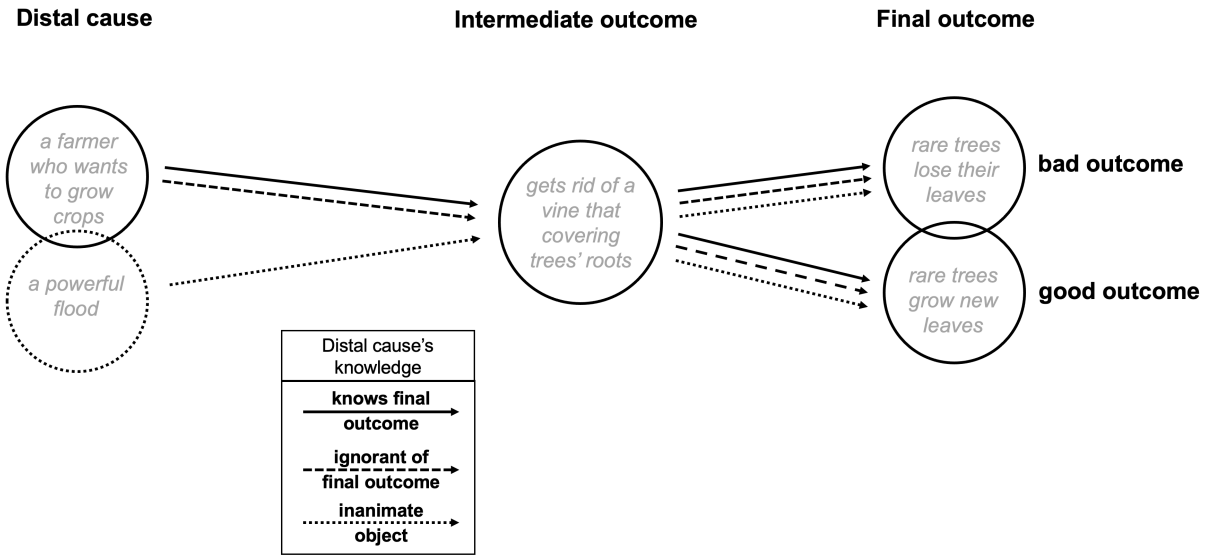


Figure 1: Study design for Experiment 1.

Methods

We report all data exclusions (if any), all manipulations, and all measures in the study.

Participants

Since we did not have a priori assumptions, we collected data from 118 participants recruited through Amazon Mechanical Turk (http://www.mturk.com) in Experiment 1. NA participants (M_{age}=36.73, SD_{age}=12.21; 47 females) answered all our demographic questions.

Materials

Participants completed 24 trials which each involved reading a brief vignette about a causal chain that was initiated by a distal cause, which led to some immediate outcome. This immediate outcome was then the more proximal cause a second, further outcome. This final outcome was either positive or negative, and the distal cause was either a knowledgeable agent (who knew that his action would lead to the occurrence of the further outcome), or an ignorant agent (who did not know that his action would lead to the occurrence of the final outcome) or an inanimate object (see Figure 1). Thus, for example, participants may have read a vignette in which an agent acted with the knowledge that the action in question would result in the occurrence of a bad outcome:

> **Knowledgeable Agent / Bad Outcome**:A farmer plans to clear a plot of land near a forest of rare trees to expand the area in which he can grow his cash crops. As he is clearing this area, an environmentalist sees him and tells him that if he clears this plot of land, he'll actually kill the rare trees in the forest by getting rid of a vine that has been protecting the trees' roots. The farmer replies that he does not care at all about the trees, he just wants to make more money by

planting cash crops. He finishes clearing the land and makes more money selling his new crops just like he planned. Not long after the vine is removed, the trees lose all their leaves.

To continue to illustrate with this example, we also altered this vignette in the Ignorant Agent conditions so that the agent simply had no way of knowing that clearing the land would lead the trees to be damaged. In the Inanimate Object conditions, we replaced the farmer with a flood that cleared the same plot of land. Finally, in the conditions where the action eventuated in a Good Outcome, the vine was described as having been damaging the tree's roots and thus removing the vine actually caused the trees to grow new leaves. Conditions were varied across scenarios.

Procedure

After reading each vignette, participants answered two questions about the events that had occurred. The first asked them to rate their agreement with a statement about the distal agent causing the outcome, as in the following example:

Causal question: The farmer caused the trees to lose all their leaves.

Participants responded to each of these questions on a scale from 1 ("Completely disagree") to 7 ("Completely agree"). The second question asked participants to complete a counterfactual question, as in the following example:

Counterfactual question: If only _____ had been different, the trees wouldn't have lost their leaves.

a. The farmer
b. The vine

After completing all 24 trials, participants were asked to complete some optional demographic questions.

Data analysis

No participants were excluded from the analyses as long as they completed the entire study (. The primary analyses were conducted with linear mixed-effects models and included random effects for both participnts and scenarios. These analyses were carried out using the the lme4 pacakage in R (Bates et al., 2014). The significance of an effect for particular factor is calculated by comparing two linear mixed-effects models that vary only in whether factor in question was included in the fixed-effects structure. to the extent that the models differ significantly in their fit, this provides evidence that the factor in question signficantly affected participants' responses.

Results

Causal judgments

We first analyzed participants' causal judgments, which revealed a main effect of the kind of *Agent* involved, $\chi^2(2) = 20.51$, $p < .001$, such that ignorant agents were overall seen as the least causal ($M = 4.99$, 95% CI = [4.71, 5.28]), even less than inanimate objects ($M = 5.42$, 95% CI = [5.08, 5.75]), $t(23) = 3.1$, $p = 0.013$. Knowledgeable agents, on the other hand, were deemed more causal than the ignorant ones ($M = 5.86$, 95% CI = [5.63, 6.1]), $t(23) = -9.12$, $p < .001$). We also observed a main effect of the kind of *Outcome* that eventuated, $\chi^2(1) = 7.12$, $p = 0.008$, suggesting that participants assigned more causality to the distal

event when the outcome was bad (M = 5.56, 95% CI = [5.3, 5.82]) than when it turned out to be good (M = 5.29, 95% CI = [5.02, 5.55]), $t(23)$ = -3.87, p = 0.001. Furthermore, these main effects were qualified by an *Agent* × *Outcome* interaction, $\chi^2(2) = 24.51$, $p < .001$.

We further decomposed this interaction and found that when agents knew about the outcome that may come as a consequence of their action ($t(23)$ = -6.63, p <.001), they were judged as much more causal for bad outcomes (M = 6.28, 95% CI = [6.03, 6.54]) than for good outcomes (M = 5.44, 95% CI = [5.16, 5.72]). In contrast, when agents were oblivious about the outcome ($t(23)$ = -1.16, p = 0.852), they were not judged to be much more causal for bad outcomes (M = 5.06, 95% CI = [4.76, 5.35]) than good outcomes (M = 4.93, 95% CI = [4.61, 5.25]). Similarly, causal judgments about non-agentic objects ($t(23)$ = 1.03, p = 0.902) did not differentiate much between bad outcomes (M = 5.35, 95% CI = [4.97, 5.72]) and good outcomes (M = 5.48, 95% CI = [5.14, 5.83]). In short, the valence of the outcome only affected participants' causal judgments when the agent at the beginning of the causal chain *knew* about the valence of the outcome (see *Figure* 2).

Counterfactual judgments

We next analyzed participants' counterfactul judgments using generalized linear mixed-effects models. These analyses again revealed a main effect of the kind of *Agent* the distal event involved, $\chi^2(2) = 15.66$, $p < 0.001$, such that ignorant agents were less frequently selected as the focus of relevant counterfactual (35%), than knowledgeable agents were (54%), z = 6.74, $p < .001$. Ignorant agents were also less likely to be selected as the counterfactual focus than objects (55%), z = -4.81, $p < .001$. In addition, we observed a main effect of *Outcome* valence once more, $\chi^2(1) = 10.51$, p = 0.001, such that the distal agent was more selected as the focus of the most relevant counterfactual for bad outcomes (47 %) than for good outcomes (57%), z = 4.74, $p < .001$. More importantly, we again observed a significant *Agent* × *Outcome* interaction effect, $\chi^2(2)$ = 22.53, $p < 0.001$.

Mirroring participants' causal judgments, we found that outcome valence strongly affected participants' counterfactual judgments when the agent was knowledgeable, such that the distal agent was the focus of counterfactuals more for bad outcomes (68%), than for good outcomes (40%), z = 7.06, $p < .001$. In contrast, when the agent was ignorant of the outcome, there was little difference in their tendency to focous on the distal agent in their counterfactual judgments in cases with bad (38%) or good outcomes (33%), z = 2.09, p = 0.292. That remains true for non-agentic objects in bad outcomes (53%) versus good outcomes (56%), z = -0.9, p = 0.946.

In short, participants' counterfactual judgments were only affected by the valence of the outcome when the agent acted with knowledge of the valence (see *Figure* 2).

Relationship between causal and counterfactual judgments

Finally, we considered the relationship between participants' causal and counterfactual judgments, and found that they were highly correlated both when considered at the level of each participants' judgments (r = 0.40, $p < 0.001$, and at the level of each the different scenarios (r = 0.67, $p < 0.001$) (see *Figure* 3). We also asked whether the counterfactual judgments mediated the observed Knowledge × Outcome interaction effect observed for ignorant agents, and found that they did: counterfactual selection mediated the relationship between causal judgment and the interaction between *Knowledge* and *Outcome* , with an average causal mediation effect (ACME) of -0.12, (95% CI = [-0.19, -0.05], $p < .001$). Controlling for the main effect of agent and knowledge, the proportion mediated is 0.29 (95% CI = [0.13, 0.49], $p < .001$).

Discussion

Experiment 1 investigated participants' causal and counterfactual judgments in a simple causal chain which eventuates in either good or bad outcomes. We found that agents who started this causal chain with

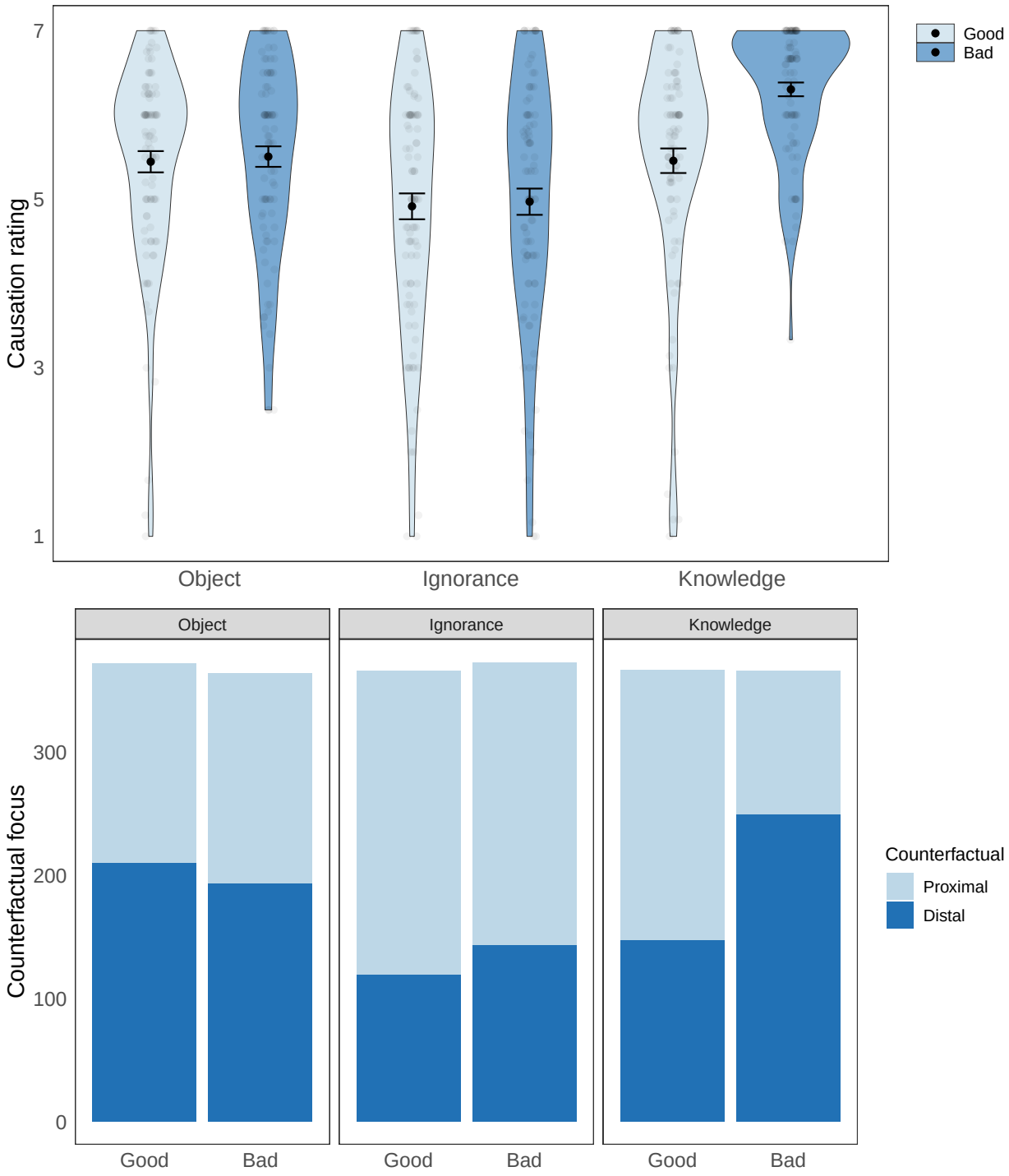


Figure 2: Participants' causal and counterfactual judgments as a function of both the kind of agent who initiated the causal chain and the valence of the outcome that eventuated. Error bars indicate +/- 1 *SEM*.

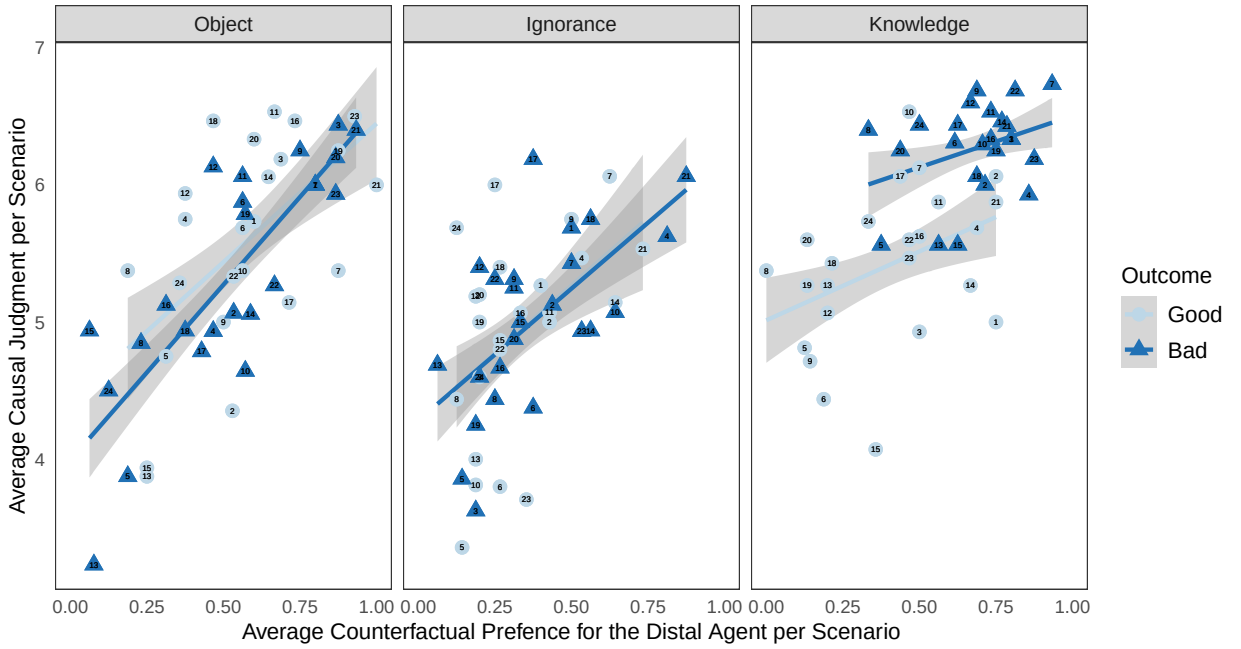


Figure 3: Depiction of the relationship between participants' causal and counterfactual judgments for each of the scenarios. Shape depicts the valence of the outcome; color indicates the kind of agent involved in the distal event; and the number indicates which of the 24 vignettes the judgments were about.

knowledge of the outcome that their action would lead to were judged to be more causal of bad than good outcomes. By comparison, agents who were ignorant about the outcome that would eventually occur were not judged to be more causal of bad (v.s. good) outcomes. Thus this first experiment demonstrates the effect of norm violations in causal chains across many scenarios and shows that it critically depends on the information available to the agent acting. The observed interaction effect between knowledge and outcome in participants' causal judgments was mirrored by a similar pattern in their counterfactual judgments: in cases where the agent *knew* about the outcome, participants were more inclined to judge that the bad (v.s. good) outcome would not have occurred if the agent had acted differently. This effect was again not seen for agents who acted in ignorance of the outcome. Across all of these conditions, we also found that participants' causal and counterfactual judgments were tightly correlated. Notably, across scenarios, we also found that the pattern of causal and counterfactual judgments for ignorant human agents resembled the pattern observed for inanimate objects, possibly suggesting a unified process for causal reasoning about both agents and inanimate objects that is underwritten by counterfactual reasoning.

Before continuing to pursue this line of reasoning, we first want to address a confound that exists in both our experiment and in much of the empirical work on the topic, specifically that the morality of the agent's action is typically confounded with the goodness or badness of the outcome (see Hitchcock & Knobe (2009) and Kominsky et al. (2015) for two notable exceptions). In the next series of studies, we independently vary the valence of the outcome and the morality of the agents' actions to allow us to ask how they separately contribute to participants' causal judgments.

Experiment 2a: Causal judgment in cases of bad outcomes and prescriptive norm violations

In this next Experiment, we began by directly addressing the confound between outcome valence and the morality of the agent's action. We did this by introducing two separate outcomes that arose from the agent's

actions (see *Fig.* 4). We used the first (primary) outcome to manipulate whether the agent's action would be seen as violating a norm, by having the agent knowingly or ignorantly acting in a way that brought about some harm. In addition, we used the secondary outcome to independently manipulate the valence of a subsequent downstream consequence of their action. The agent was always ignorant of this secondary outcome, regardless of whether it was positive or negative Critically, we still always asked whether the agent caused the downstream, secondary outcome that was either positive or negative. This design allowed us to dissociate whether or not the agent violated a norm from whether the agent acted in way that eventually led to negative or positive outcome. In this first Experiment (2a), we begin by considering cases in which the secondary outcome is always negative, and the agent could either knowingly or ignorantly bring about some other harm (the primary outcome).

A second innovation in this Experiment was to introduce a second kind of prescriptive norm violation, namely, violations of norms of rationality. Moral norm violations often involve an agent bringing about an outcome that is not in someone else's best interest, while rational norm violations often involve agents bringing about outcomes that are not in their own best interest. In this study, we also systematically vary whether the agent violated a moral or rational norm and ask whether the observed effect on causal judgments is specific to moral norm violations or extends to other kinds of prescriptive norm violations. This addition should help us address a concern raised by researchers who argue that the effect of normality is driven disproportionately by the immorality of actions. [CITEs]

To summarize, in this experiment we used 24 scenarios where an agent (i.e., the distal cause) either knowingly or ignorantly performed an action that harmed either the another person or the agent her/himself. This harm serves as the primary outcome of the agent's action. At the same time, the agent's action led to a second downstream consequence (i.e., the proximal cause) which in turn led to another downstream consequence (the secondary outcome). In this experiment the secondary outcome was always negative and the agent was always unaware of it. We then asked for participants' whether the agent (the distal cause) "caused" the secondary outcome (see *Fig. 4*).

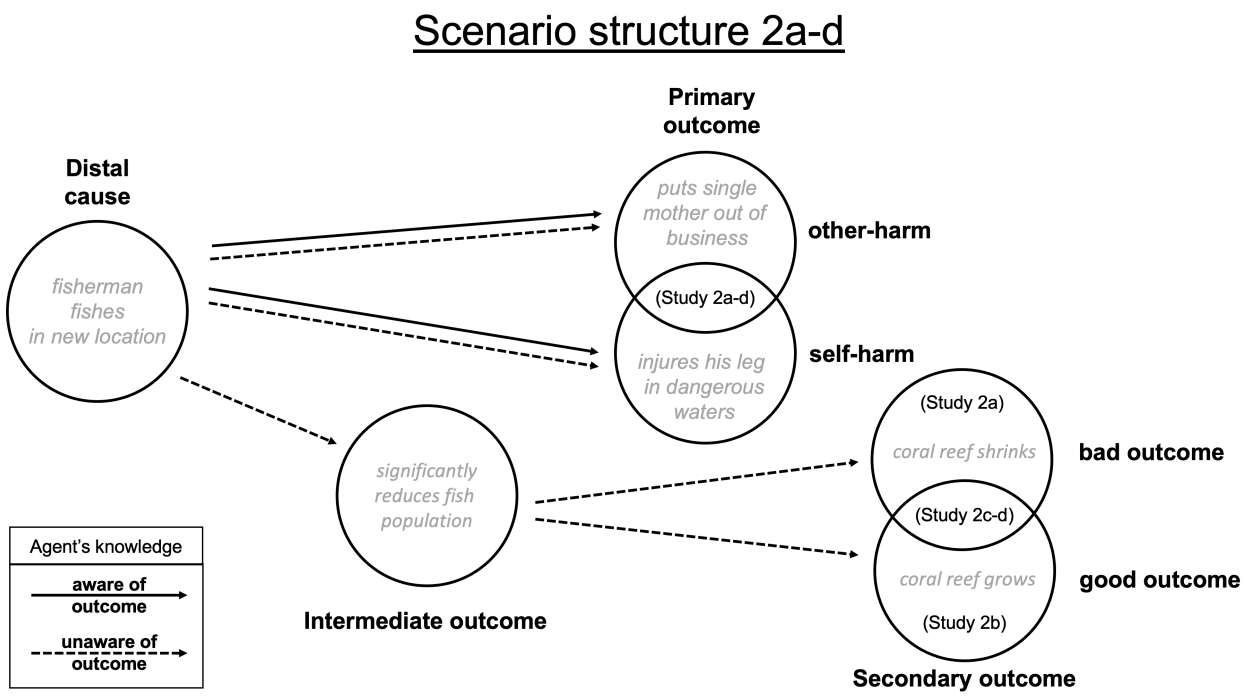


Figure 4: Study design for Experiment series 2

Methods

Participants

In Experiment 2a, 100 participants (M_{age}=34.08, SD_{age}=11.79; 36 females) were recruited through Amazon Mechanical Turk (http://www.mturk.com).

Materials

Participants completed 16 trials which each involved reading a brief vignette about a causal chain. The causal chain started with a distal agent who knowingly or ignorantly caused harm upon other people or themself. Separate from this event and unbeknownst to the agent, their action led to some immediate outcome that eventually resulted in a second outcome. In this particular experiment, the secondary outcome was invariably negative. Consider the following example from one of the scenarios we used:

> **Scenario 8 / Other-Harm / Ignorant Agent**: Harry was a fisherman applying for a license to fish in a certain coastal area. The government said that Harry could have a license to fish in this area or in a second area. They failed to tell him that choosing to fish in the first area would cause him to put a single working mother out of business. Without this information, Harry chose the license to fish in the area he initially wanted. He fished in this area every day, significantly reducing the local fish population. As a result of the lower fish population, the coral reef along the coast shrunk by several meters in every direction.

In this specific version, the agent did not know that his action (fishing in the first location) would harm another person (putting a single mother out of work). Additionally, not knowing any side effects, the agent fished at that location, which directly resulted in the intermediate event (the reduction of the fish population) that led to the occurrence of a final negative outcome (the coral reef shrinking). In the version of a knowledgeable agent, Harry was told that fishing in the area would put a single mother out of business but decided to do so anyway.

In a different version of the same scenario, the agent did something that was not to his advantage (i.e., choosing to fish at a dangerous place rather than a safer one). Consider the following example:

> **Scenario 8 / Self-Harm / Knowledgeable Agent** Harry was a fisherman applying for a license to fish in a certain coastal area. The government said that Harry could have a license to fish in this area, but that fishing there would be very difficult due to dangerous conditions. They recommended that he accept a license to fish in a second area that was much safer. Harry chose the license to fish in the area he initially wanted. He fished in this area every day, until a large wave knocked him into the rocks and he injured his leg. As a result of the lower fish population, the coral reef along the coast shrunk by several meters in every direction.

Systematically manipulating these factors resulted in an overall 2 (Harm Type) $\times$ 2 (Agent Knowledge) $\times$ 16 (Scenario) design, that was administered in a mixed within- and between-subjects fashion, such that participants saw all 16 scenarios, and on each trial were randomly assigned to read one of the 4 different versions of that scenario.

Procedure

After reading each vignette, participants rated their agreement with a statement about the distal agent causing the outcome, as in the following example:

> *Causal question*: Harry caused the coral reef to shrink by several meters in every direction.

Participants responded to each of these questions on a scale from 1 ("Completely disagree") to 7 ("Completely agree"). After completing all 16 trials, participants were asked to complete some optional demographic questions.

Data analysis

No participants were excluded from the analyses as long as they completed the entire study. The primary analyses were conducted with linear mixed-effects models and included random intercepts for both participants and scenarios as well as a random slope that measures how the impact of knowledge, harm type and their interaction may vary across scenarios.

Results

We analyzed participants' causal judgments, which revealed a main effect of *Knowledge*, $\chi^2(1) = 25.22$, $p < .001$, such that agents knowingly bringing about harm were seen as more causal ($M = 4.95$, 95% CI = [4.52, 5.38]) than those who did so ignorantly ($M = 4.06$, 95% CI = [3.68, 4.43]) regardless of the victim, $t(15.05) = -7.60$, $p < .001$, because we observed neither a main effect of *Harm Type*, $\chi^2(1) = 0.02$, $p = 0.902$, nor a *Knowledge* × *Harm Type* interaction effect, $\chi^2(1) = 0.07$, $p = 0.795$. In other words, causal judgments did not differ much between harm to self ($M = 4.49$, 95% CI = [4.07, 4.91]) and harm to others ($M = 4.51$, 95% CI = [4.08, 4.95]), $t(15.00) = 0.12$, $p = 0.907$ when the secondary outcome was negative. Taken together, findings suggest that other-harm and self-harm do not differentiate in regard to impact on causal judgments (*Fig.* 5).

Discussion

Experiment 2a revealed three insights that have not been widely documented in existing literature: first, agents who knowingly preformed actions that were detrimental to themselves or to others were judged as more causal of the subsequent negative outcome, even when the immorality or irrationality of such action was completely independent of the eventual secondary outcome. Second, the normality effect on causal judgments depends on the knowledge that one's action is violating a norm (in our paradigm, knowledge that one's action is causing harm), and not knowledge about valence of secondary outcomes: Even though the agent was always ignorant of the negative outcome we asked about (the secondary outcome), they were judged as more more of a cause of that outcome when they acted in a way that knowingly brought about some harm. Third, the effect. of normality is not specific to moral norm violations, as we found a remarkably similar pattern when the agent knowingly acted in a way that brought about harm to themselves rather than someone else. This final finding suggests that the underlying mechanism for for the norm effect in causal selection generalizes to various kinds of prescriptive norm violations.

In all of the contexts used thus far, the eventual outcome about which participants are making causal judgments is negatively valenced. Thus, one possibility is that the patterns described above depend on the negative valence of the eventual outcome, and may not extend to cases where the eventual outcome is neutral or actually positive. Such an difference might be predicted by accounts where these causal judgments are reflecting some form of motivated cognition (Alicke, Rose & Bloom, 2011; CITE). We explore this question next.

Experiment 2b: Causal judgment in cases of good outcomes and prescriptive norm violations

In Experiment 2a, we started teasing apart primary and secondary outcomes and we held fixed the secondary outcome to be negative for simplicity. However, a natural question to ask is whether the valence of the

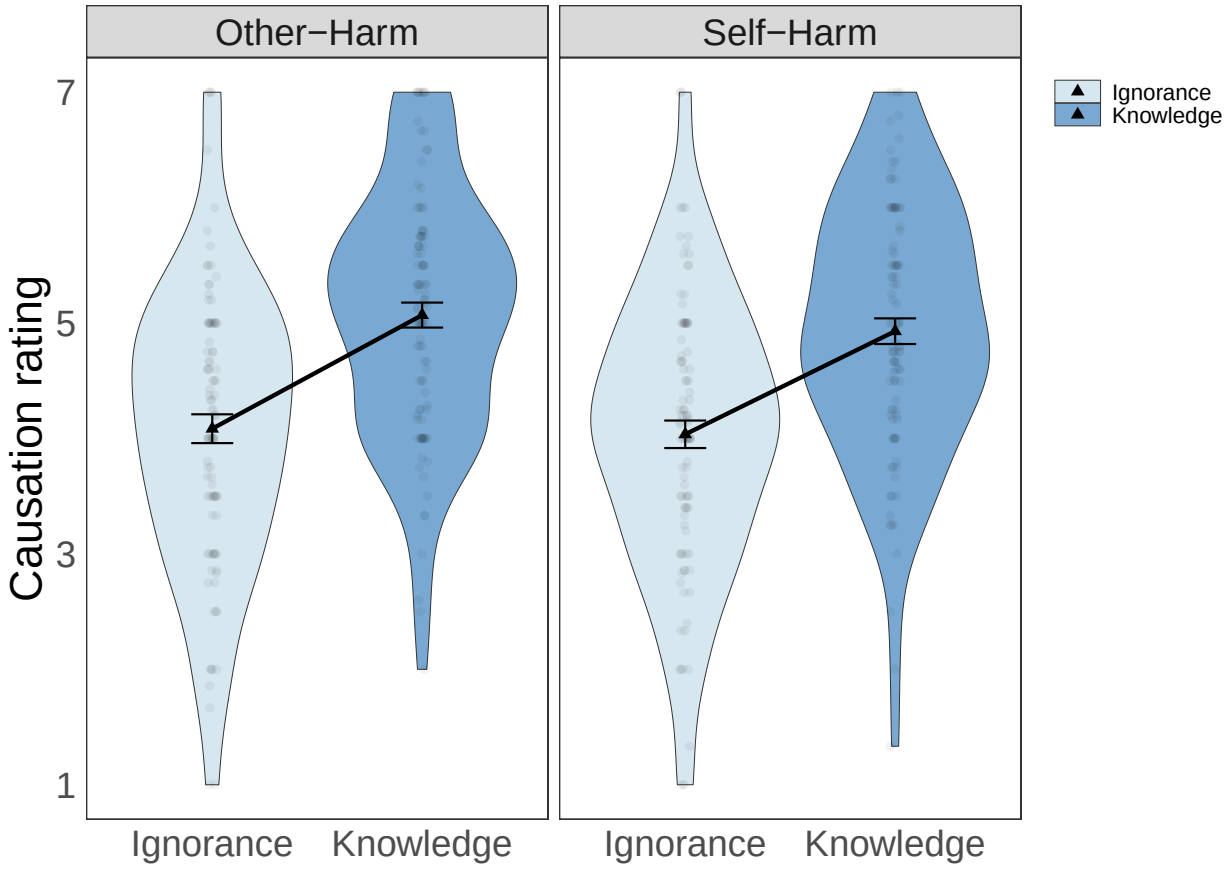


Figure 5: Participants' causal judgments as a function of both the knowledge of the agent who initiated the causal chain and the kind of norm that governed the agent's action, for both bad outcomes. Error bars indicate +/- 1 *SEM*.

outcome may have some additional effect on causal judgments, either alone or in combination with the other factors we previously manipulated. For example, it may be that people attribute more causality overall to negative than positive outcomes, or it may be that an agents' knowledge of violating a norm does not affect causal judgments when they concern positive rather than negative outcomes. To investigate this, Experiment 2b differs from Experiment 2a in only one way: the secondary outcome resulted from the intermediate event was always positive. Again, we always asked for participants' causal judgments about this *secondary outcome*. Some previous work has found an effect of moral norm violations even in good outcomes (e.g., Knobe & Hitchcock, XX; CITE), while others argue that this effect is attenuated or even nonexistent in the face of positive outcomes (e.g., Alicke, Rose & Bloom, 2011).

Methods

Participants

In Experiment 2b, 101 participants (M_{age}=33.38, SD_{age}=9.68; 50 females) were recruited through Amazon Mechanical Turk (http://www.mturk.com).

Materials

Participants completed 16 trials, each of which involved reading a brief vignette about two causal chains. As in the previous study, the agent acted, with or without knowledge, in a way that adversely impacted someone else or themself. Simultaneously, the action led to some intermediate outcome that in the end resulted in a second, final outcome that was positive, of which the distal agent was completely ignorant. Consider the following variation from a scenario we used:

> **Scenario 8 / Rational Norm / Knowledgeable Agent** Harry was a fisherman applying for a license to fish in a certain coastal area. The government said that Harry could have a license to fish in this area, but that fishing there would be very difficult due to dangerous conditions. They recommended that he accept a license to fish in a second area that was much safer. Harry chose the license to fish in the area he initially wanted. He fished in this area every day, until a large wave knocked him into the rocks and he injured his leg. As a result of the lower fish population, the coral reef along the coast grew by several meters in every direction.

This design, again, resulted in an overall 2 (Harm Type) × 2 (Agent Knowledge) × 16 (Scenario) design, that was administered in a mixed within- and between-subjects fashion, such that participants saw all 16 scenarios, and on each trial were randomly assigned to read one of the 4 different versions of that scenario.

Procedure

After reading each vignette, participants rated their agreement with a statement about the distal agent causing the outcome, as in the following example:

> *Causal question*: Harry caused the coral reef to grow by several meters in every direction.

Participants responded to each of these questions on a scale from 1 ("Completely disagree") to 7 ("Completely agree"). After completing all 16 trials, participants were asked to complete some optional demographic questions.

Data analysis

No participants were excluded from the analyses as long as they completed the entire study. The primary analyses were conducted with linear mixed-effects models and included random intercepts for both participants and scenarios as well as a random slope that measures how the impact of knowledge, harm type and their interaction may vary across scenarios.

Results

We analyzed participants' causal judgments, which did not reveal any main effects or interaction, including agents' *Knowledge* about the direct consequence of their action, $\chi^2(1) = 2.69$, $p = 0.101$, the *Harm Type*, $\chi^2(1) = 0.11$, $p = 0.744$, and the *Knowledge* × *Harm Type* interaction $\chi^2(1) = 0.83$, $p = 0.362$ (*Fig.* 6). That is, when the final outcome was good, knowledgeable agents were not deemed more causal ($M = 4.74$, 95% CI = [4.39, 5.10]) than ignorant ones ($M = 4.63$, 95% CI = [4.25, 5.01]), $t(13.42) = -1.56$, $p = 0.141$, no matter if the harm incurred at the same time affected the agents themselves ($M = 4.72$, 95% CI = [4.33, 5.10]) or others ($M = 4.66$, 95% CI = [4.27, 5.04]), $t(13.96) = -0.47$, $p = 0.649$.

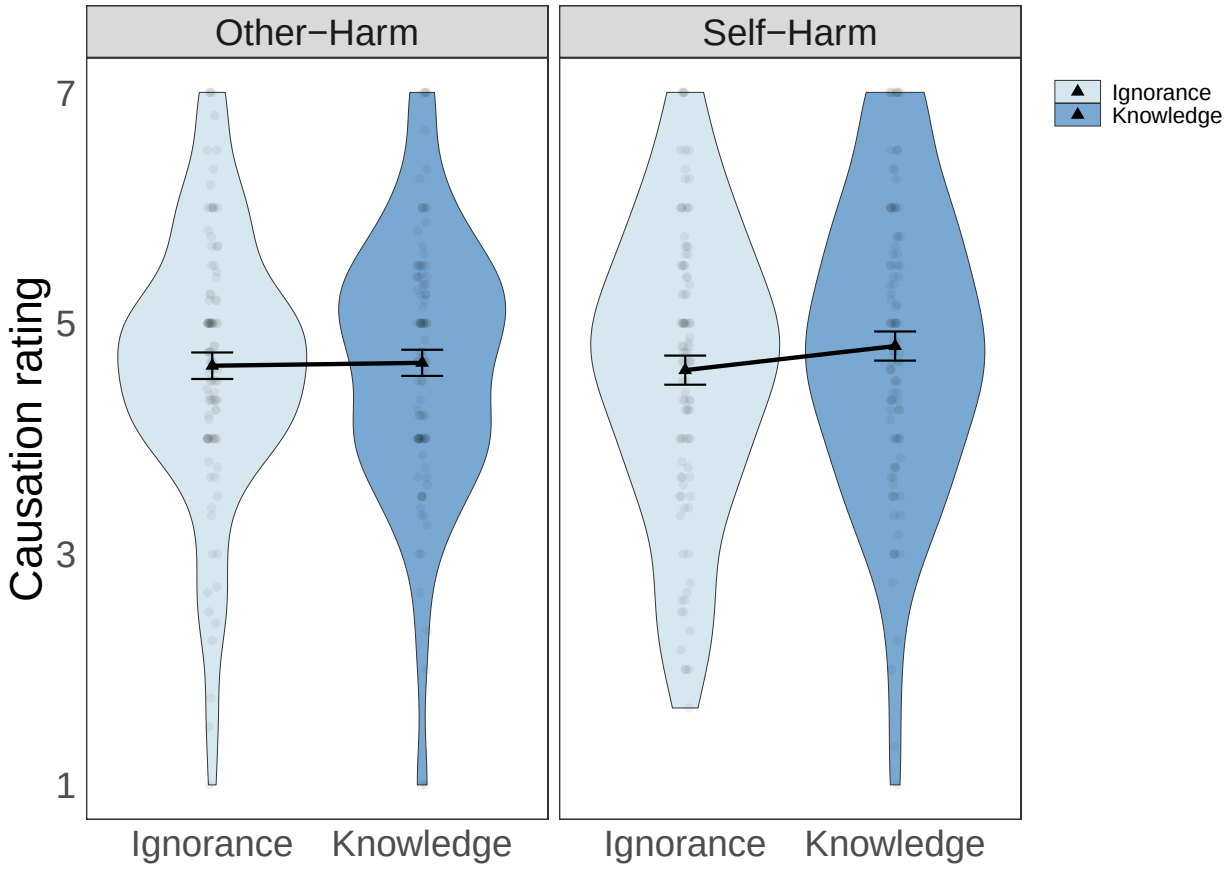


Figure 6: Participants' causal judgments as a function of both the knowledge of the agent who initiated the causal chain and the kind of norm that governed the agent's action, for good outcomes. Error bars indicate +/- 1 *SEM*.

Discussion

In contrast to the findings in Experiment 2a, in Experiment 2b we observed much weaker and non-significant effects of agents' knowledge that they were violating a norm. That is, agents who knew that they were violating a norm when acting were judged to be somewhat similarly causal of downstream, secondary outcomes as agents who were ignorant that they were violating a norm when acting. This pattern broadly held for both rational and moral norms. Taken together then, Experiment 2a and 2b suggest that the valence of the secondary outcome may influence the extent to which agents' knowledge impacts causal judgments.

At the same time, it is difficult to estimate such an interaction effect, because these two studies were conducted independently. Accordingly, a better way to directly investigate such a potential interaction would be to vary the valence of the outcome in addition to the other key factors manipulated in Experiments 2a-b. We pursue this next.

Experiment 2c: Causal judgment in cases of bad/good outcomes and prescriptive norm violations

In Experiment 2a and 2b, we investigated causal cognition in bad and good outcomes separately. Here, we replicate these studies in a within-subjects design by systematically manipulating the valence of the secondary outcome alongside the knowledge of the agent and the kind of normality at issue in the scenario (see *Fig.* 4). Based on the prior results, we specifically want to investigate whether we observe an interaction effect, whereby the impact of the agent's knowledge on causal judgments is larger when the secondary, downstream consequence is negative rather than positive. We again predict that the pattern will be quite similar when the agent violates a moral vs. rational norm, and thus do not anticipate a three-way interaction.

Methods

Participants

In Experiment 2c, 214 participants (M_{age}=33.18, SD_{age}=8.98; 94 females) were recruited through Amazon Mechanical Turk (http://www.mturk.com).

Materials

Participants completed 16 trials which each involved reading a brief vignette about events in two related causal chains. The general design remains the same as that in Experiment 2a and 2b except that the secondary outcome resulting from the intermediate outcome of the agent's action was either negative or positive.

This design resulted in an overall 2 (Harm Type) $\times$ 2 (Agent Knowledge) $\times$ 2 (Outcome Valence) $\times$ 16 (Scenario) design, that was administered in a mixed within- and between-subjects fashion, such that participants saw all 16 scenarios, and on each trial were randomly assigned to read one of the 8 different versions of that scenario.

Procedure

After reading each vignette, participants rated their agreement with a statement about the distal agent causing the final secondary outcome as before by responding on a scale of 1 ("Completely disagree") to 7 ("Completely agree"), and were asked to complete some optional demographic questions after rating all 16 scenarios.

Data analysis

No participants were excluded from the analyses as long as they completed the entire study. The primary analyses were conducted with linear mixed-effects models and included random intercepts for both participants and scenarios as well as a random slope that measures how the impact of knowledge, outcome and their interaction may vary across scenarios.

Results

We analyzed participants' causal judgments. To start with, we found a significant main effect in *Knowledge*, $\chi^2(1) = 13.50$, $p < .001$, where knowledgeable agents were deemed more causal ($M = 5.06$, 95% CI = [4.81, 5.32]) than ignorant agents ($M = 4.65$, 95% CI = [4.33, 4.97]) for the secondary outcome, $t(15.12) = -6.45$, $p < .001$. No significant results came from analysis on main effects of *Harm Type*, $\chi^2(1) = 0.02$, $p = 0.879$ or the valence of *Outcome*, $\chi^2(1) = 0.01$, $p = 0.907$. Moreover, we saw a significant *Knowledge* × *Outcome* interaction effect that ties back to the results of Study 2a and 2b, $\chi^2(1) = 9.38$, $p = 0.002$, such that when secondary outcomes were bad, participants attributed more causation to knowledgeable agents ($M = 5.20$, 95% CI = [4.90, 5.50]) than ignorant agents ($M = 4.58$, 95% CI = [4.22, 4.95]), $t(15.17) = -6.50$, $p < .001$. But when secondary outcomes turned out to be good, the causal ratings differed less between agents who caused harm knowingly ($M = 4.93$, 95% CI = [4.66, 5.20]) and those who did so obliviously ($M = 4.72$, 95% CI = [4.42, 5.01]), $t(15.06) = -2.73$, $p = 0.066$ (*Fig.* 7). There was no significant *Knowledge* × *Harm Type* × *Outcome* interaction effect, $\chi^2(1) = 0.24$, $p = 0.625$, *Harm Type* × *Outcome* interaction effect, $\chi^2(1) = 3.17$, $p = 0.075$, or *Harm Type* × *Knowledge* interaction effect, $\chi^2(1) = 0$, $p = 0.963$.

Discussion

These results revealed a set of key findings. First, we again replicated the effect of knowledge. When agents do an action with knowledge that it will harm themselves or someone else, they are seen as more of a cause of unrelated and unforeseen downstream consequences of their actions. Second, these results also demonstrate the expected interaction effect between knowledge and the valence of the secondary outcome. Specifically, an agent's knowledge had a bigger impact on causal judgments when the unforeseen downstream consequence was negative rather than positive. This interaction occurs in both moral norm violations (i.e., harming another) and rational norm violations (i.e., harming oneself), and thus provides further evidence that judgments of causal selection are working in a similar manner across different types of prescriptive norms. This fact makes it unlikely that the overall pattern could be explained by appealing to polysemy or the motivation to judge someone harshly because she/he did something that harmed another person.

In Study 1, we found that the impact of normality on participants' causal judgments could be explained by differences in their counterfactual reasoning. Then, in Studies 2a-2c, we found that that the effect of normality depended critically on the agent's knowledge of violating a norm and that this effect interacted with the valence of the downstream consequence participants made a causal judgment about. We next want to ask whether the specific effects observed in Studies 2a-c can also be explained by changes in participants' counterfactual reasoning. In addition, we wanted to collect ratings of the morality and rationality of the agents' actions in all conditions.

Experiment 2d: Counterfactual judgments and morality/rationality ratings in cases of bad/good outcomes

Building on Experiments 2a-2c, Experiment 2d asked participants to make counterfactual selection judgments (similar to Experiment 1), as a way of further investigating whether these effects can be explained by theories that appeal to counterfactual reasoning. In addition, participants were asked to rate the morality and rationality of the agent's action. These ratings serve two purposes. First, they serve as a manipulation

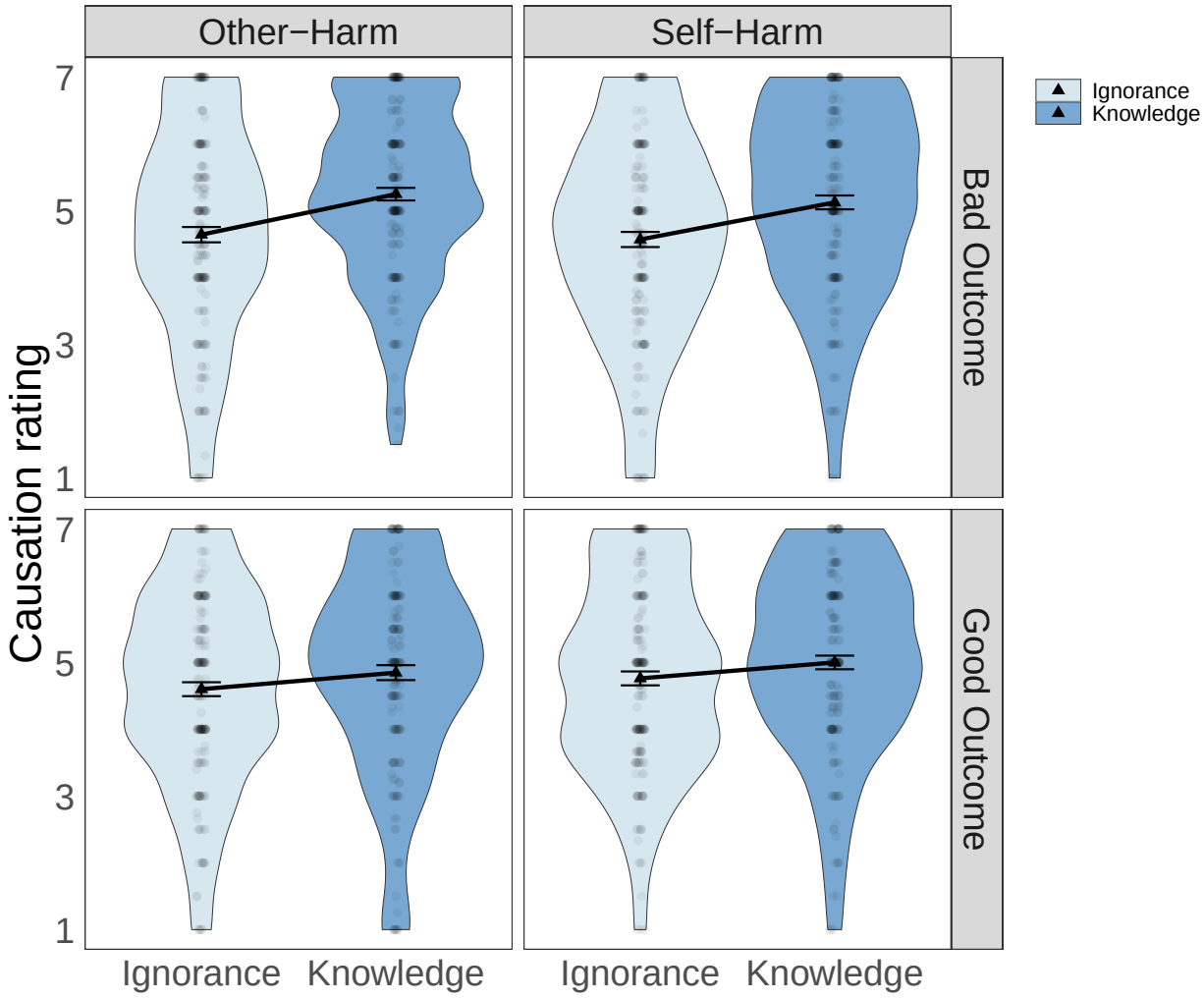


Figure 7: Participants' causal judgments as a function of both the knowledge of the agent who initiated the causal chain and the kind of norm that governed the agent's action, for both bad outcomes (left) and good outcomes (right). Error bars indicate +/- 1 *SEM*.

check, allowing us to verify, for example, that our manipulation of the agent's knowledge of violating a norm impacted participants' perceptions of the morality or rationality of the agent's action. Second, combining the data from this Experiment with the data from Experiments 2a-c, we can take advantage of the naturally occurring variance in perceived immorality or irrationality across the large number of contexts we used. That is, instead of analyzing all the contexts with simple binary condition labels (e.g., knowledgeable, moral norm), we can ask whether causal judgments are predicted by the continuous measures of perceived immorality and irrationality.

Methods

Participants

For ratings of the morality and rationality of the agent's actions, 199 participants (M_{age}=35.19, SD_{age}=11.16; 89 females) were recruited. For ratings of which counterfactual choices were relevant, 203 participants (M_{age}=33.42, SD_{age}=10.58; 99 females) were recruited. All participants were recruited through Amazon Mechanical Turk (http://www.mturk.com).

Materials

In both studies, participants completed 16 trials which each involved reading the brief vignettes about causal chains as in the previous studies. As in Study 2c, the design for both studies was a 2 (Harm Type) $\times$ 2 (Agent Knowledge) $\times$ 2 (Outcome Valence) $\times$ 16 (Scenario) design that was administered in a mixed within- and between-subjects fashion, such that participants saw all 16 scenarios and on each trial were randomly assigned to read 1 of the 8 different versions of that scenario.

Procedure

For the counterfactual question, participants selected the best way to complete a counterfactual statement about the prevention of the outcome from two options. For example, in the good outcome versions of the scenario with Harry, this question reads as follows:

> *Counterfactual Question*: If only _____ had been different, the coral reef would not have grown by several meters in every direction.
> a. Harry
> b. the fish population

For the ratings of the morality and rationality, participants first read the brief vignette and then answered two questions about the morality and rationality of the agent's action. In the example scenario we have been using throughout, these questions read as follows:

> *Morality Question*: Was it immoral for Harry to choose the lisence to fish in the area he initially wanted?

> *Rationality Question*: Was it irrational for Harry to choose the lisence to fish in the area he initially wanted?

Participants answered both questions on a 7-point Likert scale from 1 ('Not at all') to 7 ('Completely'), with a midpoint of 4 ('In between').

Participants were asked to complete a brief demographic questionnaire after completing all 16 trials.

Data analysis

No participant was excluded from the analyses as long as the entire study was completed. For analyses involving causal judgments, we used data collected in Studies 2a-2c. We then combined and analyzed all relevant data at the level of the various scenarios. Besides...

Results

First, we analyzed whether participants' judgments of the morality and rationality of the agents' actions tracked our manipulations as intended. We then go on to investigate the counterfactual selection judgments.

Morality Question

Note that we report all analyses conducted for this experiment to ensure comprehensive coverage, including results that are not directly relevant to our primary predictions.

We started with participants' moral judgments and found main effects across *Knowledge*, *Harm Type* and *Outcome* respectively. First, the main effect of the agent's *Knowledge* about the consequence of their action, $\chi^2(1) = 0$, $p < .001$, indicates that events knowledgeable agents were overall seen as less moral ($M = 4.03$, 95% CI = [3.73, 4.34]) than ignorant agents ($M = 5.62$, 95% CI = [5.37, 5.88]). Second, morality judgments also differentiated between *Harm Types*, $\chi^2(1) = 28.18$, $p < .001$, which suggests that agents who caused harm to others were considered less moral ($M = 4.24$, 95% CI = [3.99, 4.48]) than those who did harm to themselves ($M = 5.42$, 95% CI = [5.07, 5.77]). Third, a main effect of the final *Outcome* was discovered, $\chi^2(1) = 23.145$, $p < .001$: participants considered agents whose actions resulted in a bad outcome to be more immoral ($M = 4.46$, 95% CI = [4.14, 4.78]) than agents whose actions resulted in a good one ($M = 5.19$, 95% CI = [4.95, 5.44]).

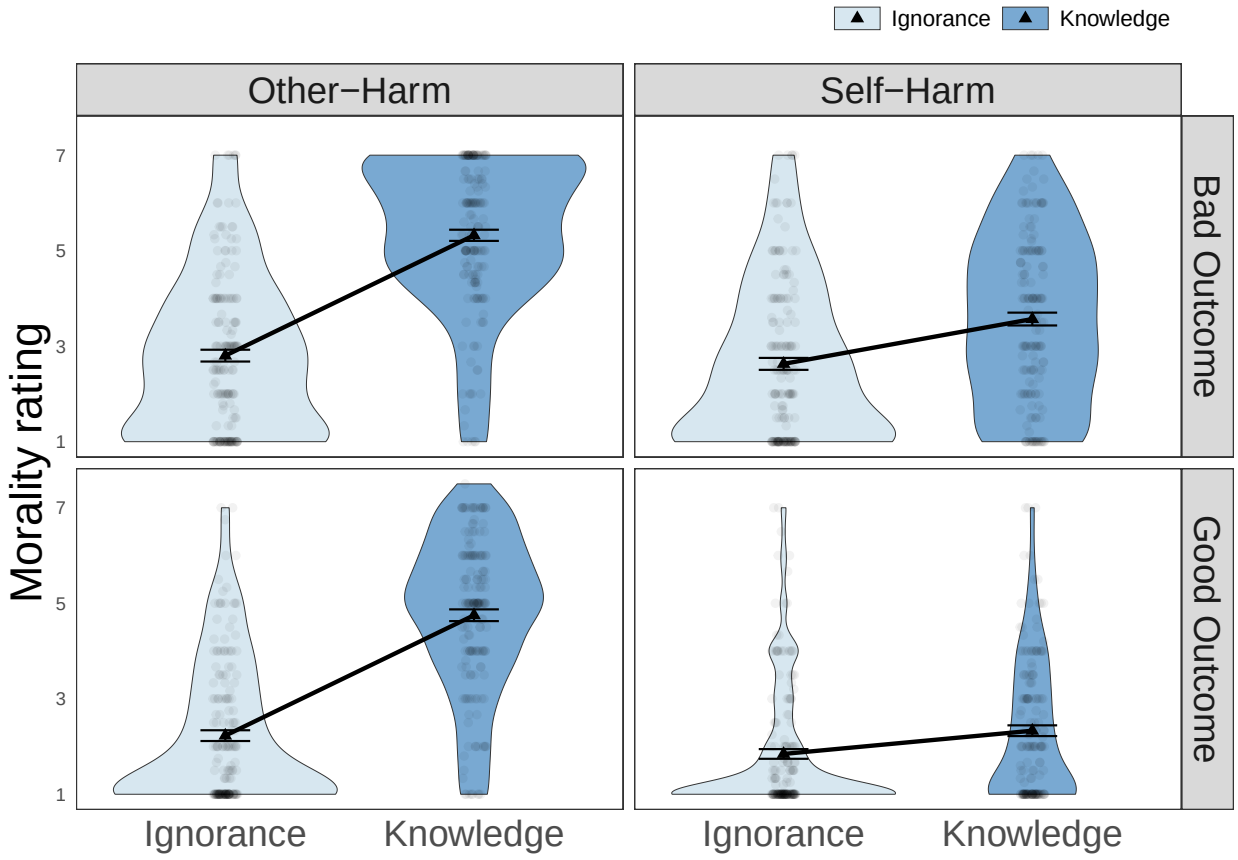
We then looked into interaction effects for moral ratings. A significant *Harm Type* × *Knowledge* interaction $\chi^2(1) = 260.1$, $p < .001$, was observed. More specifically, when agents acted ignorantly, moral judgment on them harming selves ($M = 5.75$, 95% CI = [5.41, 6.08) and that on harming others ($M = 5.5$, 95% CI = [5.24, 5.77) was not outstanding, but when acting knowingly, agents who harmed others were deemed much more immoral ($M = 2.97$, 95% CI = [2.7, 3.25) than those who merely acted to their own disadvantage ($M = 5.09$, 95% CI = [4.69, 5.49). There was also a (unexpectedly??) significant *Harm Type* × *Outcome* interaction, $\chi^2(1) = 17.894$, $p < .001$, such that when the final outcome was good ($t(20) = -9.47$, $p = 0$), participants believed agents who committed harm on others were less moral ($M = 4.48$, 95% CI = [4.25, 4.72]) than those who harmed only themselves ($M = 5.9$, 95% CI = [5.57, 6.24), but when this secondary outcome was bad ($t(20) = -6.27$, $p = 0$), this difference in morality rating between other-harming agents ($M = 3.99$, 95% CI = [3.69, 4.3) and self-harming agents became smaller ($M = 4.93$, 95% CI = [4.54, 5.33). Analyses did not yield statistically significant results in the *Knowledge* × *Outcome* interaction, $\chi^2(1) = 2.95$, $p = 0.086$ or the three-way *Knowledge* × *Harm Type* × *Outcome* interaction, $\chi^2(1) = 2.05$, $p = 0.152$.

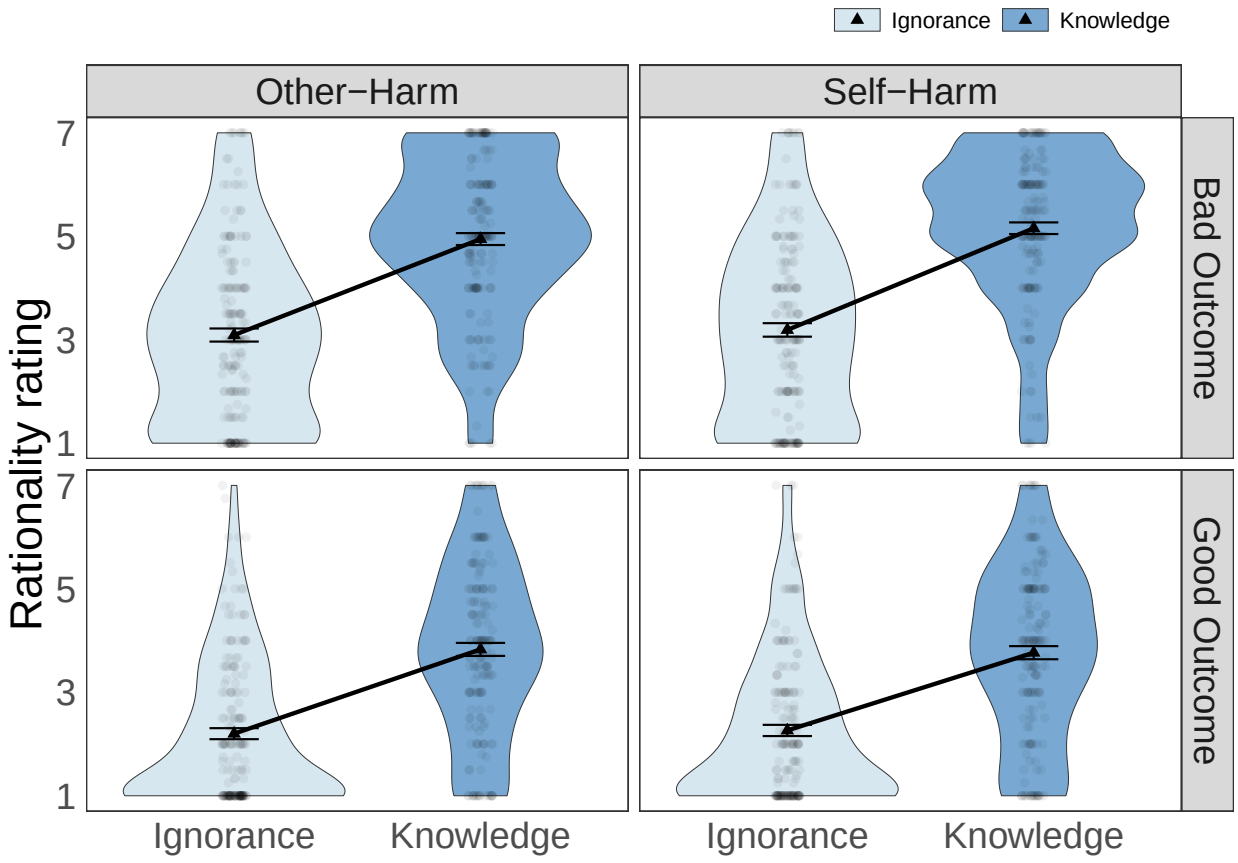
Rationality Question

We next examined participants' judgments of rationality in the same manner. We found main effects in *Knowledge* and *Outcome*, but not *Harm Type*, $\chi^2(1) = 0.77$, $p = 0.381$. The *Knowledge* effect, $\chi^2(1) = 40.867$, $p < .001$, indicates that agents who knowingly caused harm were believed to be less rational ($M = 3.56$, 95% CI = [3.29, 3.83]) than agents who ignorantly did so ($M = 5.32$, 95% CI = [5.1, 5.54]). We also discovered a main effect of the final *Outcome*, $\chi^2(1) = 24.11$, $p < .001$: agents whose action resulted in a bad outcome were considered less rational ($M = 3.91$, 95% CI = [3.64, 4.17]) than agents whose action resulted in a good one ($M = 4.97$, 95% CI = [4.73, 5.22]).

Like in the morality rating, responses to the rationality question saw a (unexpected??) *Harm Type* × *Outcome* interaction, $\chi^2(1) = 5.52$, $p = 0.019$, such that when the agent's action eventually led to a bad

outcome ($t(19) = 1.59$, $p = 0.41$), those who did not act in alignment with their own best interest were thought to be less rational ($M = 3.77$, 95% CI = [3.45, 4.08]) than those who harmed others ($M = 4.05$, 95% CI = [3.72, 4.37]), but when the end outcome was positive ($t(19) = 0.02$, $p = 1$), participants did not differ much in their rationality ratings about harming others ($M = 4.98$, 95% CI = [4.67, 5.28) and those about harming selves ($M = 4.97$, 95% CI = [4.66, 5.28). Moreover, we found a *Knowledge* × *Outcome* interaction effect, $\chi^2(1) = 8.89$, $p = 0.003$, which shows that when agents acted with knowledge ($t(20) = -7.85$, $p = 0$), the ones who indirectly brought about good outcomes ($M = 4.18$, 95% CI = [3.85, 4.52]) were seen as more rational than those who brought about bad outcomes ($M = 2.94$, 95% CI = [2.65, 3.23]). However, when agents acted ignorantly ($t(20) = -5.62$, $p = 0$), rationality judgments differed less between those who eventually caused bad outcomes ($M = 4.87$, 95% CI = [4.56, 5.19]) and those who caused good outcomes ($M = 5.76$, 95% CI = [5.54, 5.99]). There was no significant *Knowledge* × *Harm Type* × *Outcome* interaction effect, $\chi^2(1) = 2.15$, $p = 0.143$, nor *Harm Type* × *Knowledge* interaction effect, $\chi^2(1) = 0.21$, $p = 0.651$.





Counterfactual Question

We asked participants about the relevance of counterfactuals by examining their choice of the cause - the human agent or the external environment. Then we analyzed responses using generalized linear mixed-effects models and found main effects of *Harm Type*, $\chi^2(1) = 5.51$, $p = 0.019$, and agent *Knowledge*, $\chi^2(1) = 17.09$, $p < .001$. In other words, agents were more likely to be selected as the counterfactual focus when the harm affected others (44%) instead of themselves (39%), z = -7.62, $p < .001$, and when the agent knew about the consequence (50%) instead of being oblivious (33%), z = -7.62, $p < .001$ (see *Figure* 8). There was not a significant main effect of *Outcome*, $\chi^2(1) = 2.92$, $p = 0.087$.

As for interactions, we once again saw a significant *Knowledge* × *Outcome* effect, $\chi^2(1) = 7.03$, $p = 0.008$, suggesting that participants' counterfactual choice tends to differentiate more between knowledgeable agent and ignorant agent when the event resulted in a bad outcome, z = -6.98, $p < .001$, than in a good outcome, z = -4.35, $p < .001$. Specifically, when knowledge about action consequence was available, the agent was more likely to be selected as the counterfactual focus than the external environment for bad outcomes (57%), than for good outcomes (43%), $z = 3.85$, $p < .001$. In contrast, when agents were oblivious, there was little difference in participants' counterfactual choice between bad outcomes (33%) and good outcomes (33%), $z = 0.1$, $p = 1$. Other than that, analyses did not reveal significant results in the *Knowledge* × *Harm Type* interaction, $\chi^2(1) = 0.41$, $p = 0.521$, the *Harm Type* × *Outomce* interaction, $\chi^2(1) = 1.23$, $p = 0.267$, or the three-way *Knowledge* × *Harm Type* × *Outcome* interaction, $\chi^2(1) = 0$, $p = 1$. Like we see in Experiment 1, counterfactual choice only differentiated between good and bad outcomes when knowledge was available to the agents (see *Figure* 8).

Combined Analyses

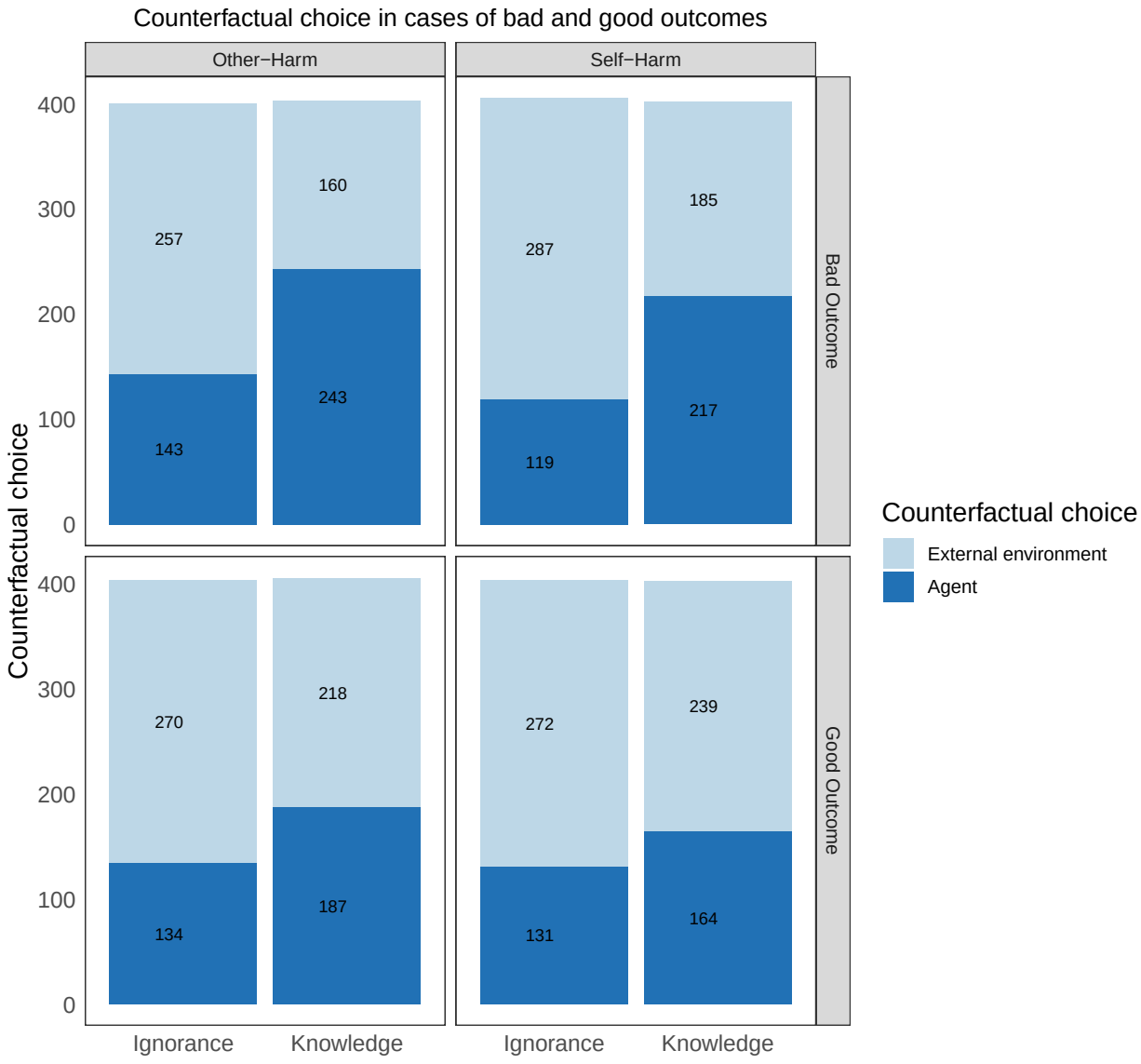


Figure 8: Counterfactual choices in good and bad outcomes.

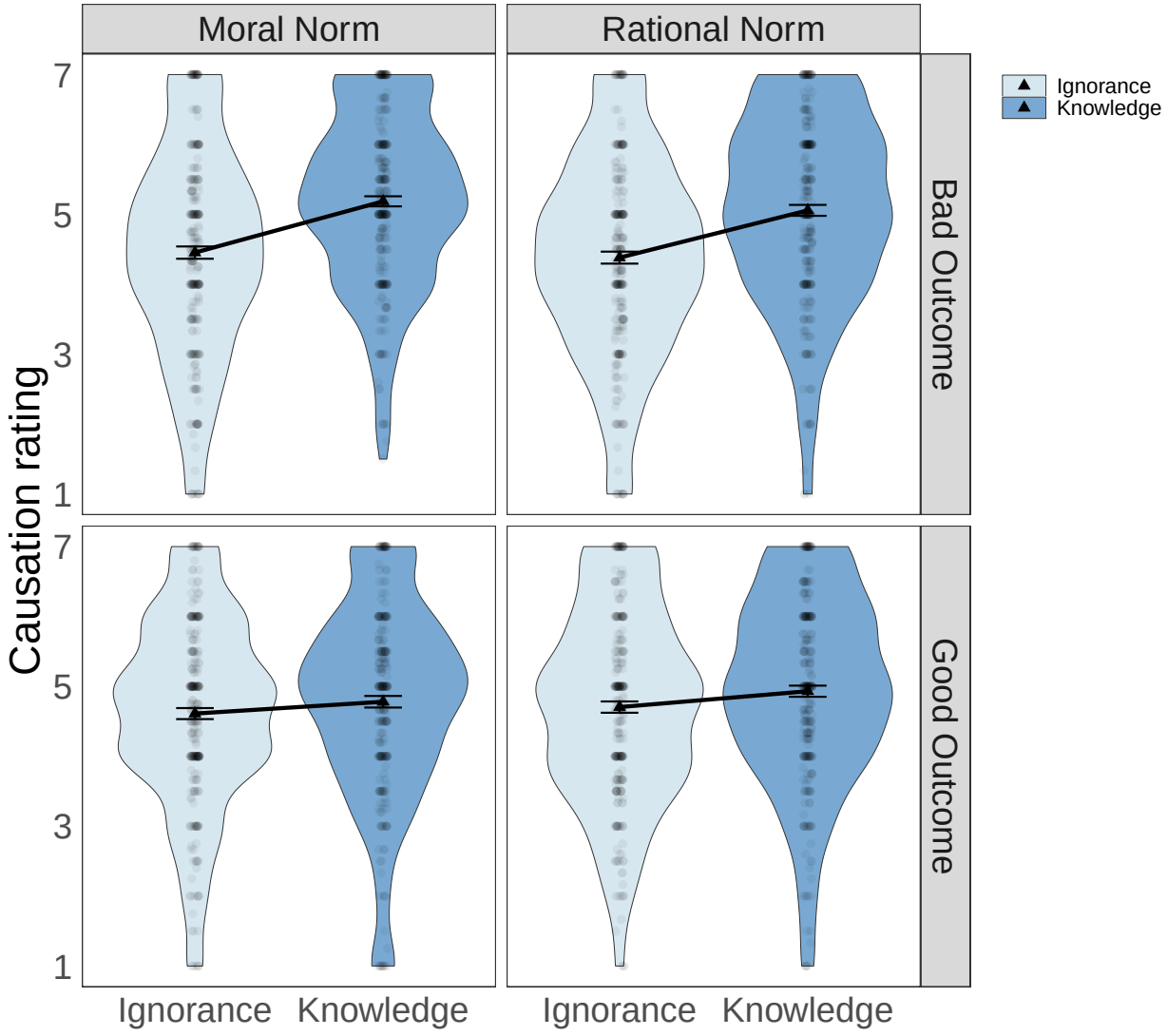


Figure 9: Participants' causal judgments as a function of knowledge status, outcome valence, and the kind of harm caused. Error bars indicate +/- 1 *SEM*.

Pooled causal judgments Combining and analyzing causal judgment data from Experiments 2a-2c, we discovered a significant main effect in in *Knowledge*, $\chi^2(1) = 11.07$, $p < .001$ as well as a significant interaction effect between *Knowledge* × *Outcome*, $\chi^2(1) = 17.89$, $p < .001$. In other words, knowledgeable agents were considered more causal (M = 4.96, 95% CI = [4.67, 5.24]) than ignorant agents (M = 4.5, 95% CI = [4.19, 4.81]) for the secondary outcome, $t(14.97)$ = -8.4, $p < .001$, and the magnitude of this knowledge effect depends on the outcome valence, such that when secondary outcomes were bad, more causation was assigned to knowledgeable agents (M = 5.12, 95% CI = [4.78, 5.46]) than ignorant agents (M = 4.37, 95% CI = [4.03, 4.71]), $t(15.07)$ = -8.34, $p < .001$, while in positive outcomes, the difference in causal judgment became smaller between agents who acted knowingly (M = 4.8, 95% CI = [4.52, 5.07]) and those who did so ignorantly (M = 4.62, 95% CI = [4.32, 4.93]), $t(14.64)$ = -2.91, $p = 0.048$ (*Fig.* 9). These results align with our findings in Experiment 2c.

Relationship between causal and counterfactual judgments We considered the relationship between causal judgments and counterfactual choices, using data from the current study and Experiments 2a-2c.

Mirroring analyses conducted in Experiment 1, we investigated the relationship between participants' causal and counterfactual judgments, and found that they were highly correlated across experimental conditions ($r = 0.52$(see *Figure* 10). We also looked into whether the counterfactual judgments mediated the observed Knowledge × Outcome interaction effect observed for ignorant agents, and found that they did: counterfactual selection mediated the relationship between causal judgment and the interaction between *Knowledge* and *Outcome* , with an average causal mediation effect (ACME) of -0.09, (95% CI = [-0.16, -0.03], $p = 0.002$). Controlling for the main effect of agent and knowledge, the proportion mediated is 0.33 (95% CI = [0.17, 0.58], $p = 0.002$).

Discussion

Experiment 2d built on prior studies by asking for additional ratings on morality and rationality of agents' actions as well as counterfactual judgments comparing the agent with external factors such as the environment.

At a broad level the morality and rationality ratings confirmed that our manipulations worked and critically we observed an effect of the agent's knowledge. For example, immorality ratings were highest when an agent knowingly brought about harm to another, and agents who knowingly harmed themselves were judged to be highly irrational but not immoral. Interestingly, we observed two unpredicted effects in ratings of rationality. First, participants rated agents who knowingly harming others to be irrational in addition to immoral, even though the action often benefited the agent in some way. Second, agents were seen as more irrational for knowingly causing harm when the secondary, unforeseen, outcome was negative rather than positive. We will return to these ratings and their relationship to causal judgments in the final set of analyses.

More critically, the judgement of counterfactual selection broadly confirmed our hypothesis. As in Experiment 1, human agents were more likely to be selected as the crux of the counterfactual choice when they were aware of the consequences of their actions and this difference was more pronounced in negative than positive outcomes. Additionally, this same pattern held if the action inflicted harm on someone else rather than oneself. More critically, across all of these conditions and scenarios, we found high correlation between participants' causal and counterfactual judgments. Taken together, Experiment 2d suggests that the strong tie between causal reasoning and counterfactual cognition applies in cases of prescriptive norm violations, and in fact, the observed interaction effect between agent knowledge and outcome valence is mediated by differences in counterfactual judgments.

While these results suggest that differences in counterfactual reasoning be able to help account for the interaction effect between the agent's knowledge of violating a norm and the valence of the secondary, unforeseen, outcome, they also a further question: Why do these two factors influence participants' counterfactual reasoning? Given the general tie between normality and counterfactual reasoning, one possibility is that participants are perceiving the negatively valenced outcomes to be more abnormal than the positively valenced outcomes, and thus are more likely to consider counterfactual alternatives in which the negative

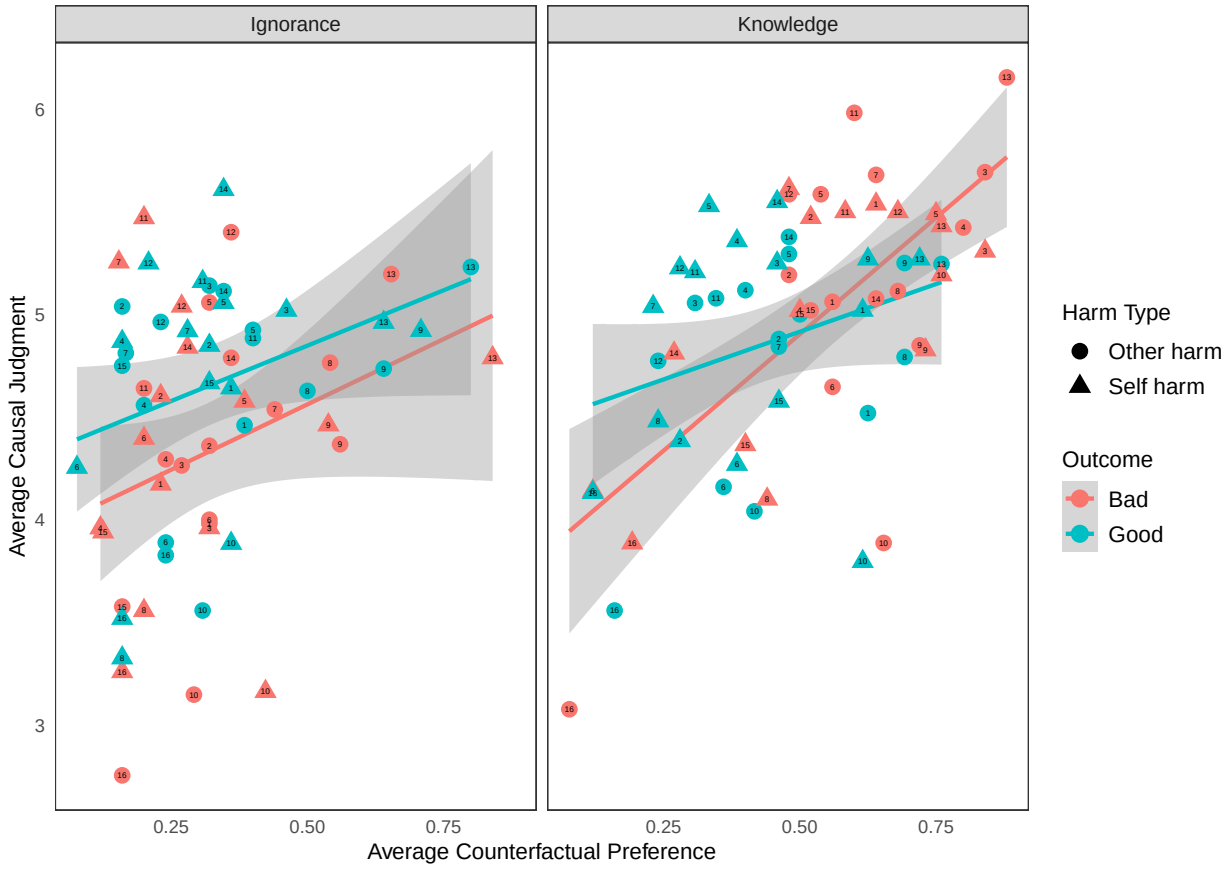


Figure 10: Depiction of the relationship between participants' causal and counterfactual judgments for each of the scenarios.

outcome would not have arisen (CITE - I think the early counterfactual literature discusses this. Maybe Neal Roese?). This difference in counterfactual reasoning may then interact with the well-documented effect of action-based norm violations on counterfactual reasoning. We next pursue this possibility (in Experiment 3a-3c) by considering whether a similar interaction effect arises when the outcomes are merely descriptively normal vs abnormal (i.e., likely or unlikely) rather than having a negative or positive valence.

Experiment 3

We previously explored the role of prescriptive information by manipulating the valence of outcomes and found that outcome valence affected causal judgments, partially through an interaction with the normality of the agent's action that led to this outcome. However, it remains unclear whether this effect is unique to prescriptive differences in the outcome or reflects a broader impact of the perceived normality of the outcome. Prior studies suggest that people's perception of normality incorporates both prescriptive and descriptive norms—what is morally ideal and what is statistically expected (Bear & Knobe, 2017). To further investigate the general role of normality, we shifted our focus from the prescriptive value of outcomes to their likelihood of occurrence. We expected to again see the interaction between the acting agent's knowledge and the probability of outcomes, which would offer support to a more general account of how normality influences causal judgment by changing participants' counterfactual reasoning.

Experiment 3a: Causal judgment in cases of probable/improbable outcomes and prescriptive norm violations

Similar to the main causal chains in Experiment 2, Experiment 3a features an agent (i.e., the distal cause) performing an action that is itself the proximal cause for a subsequent outcome which is either probable or improbable in terms of chance of occurrence. This secondary outcome was still unforeseen by the agent. The agent was at the same time knowledgeable or ignorant about a separate primary consequence that was either negative for the agents themselves or negative for other people. In this study, we asked participants to what extent they saw the agent (i.e., the distal cause) as the cause of the likely or unlikely secondary outcome.

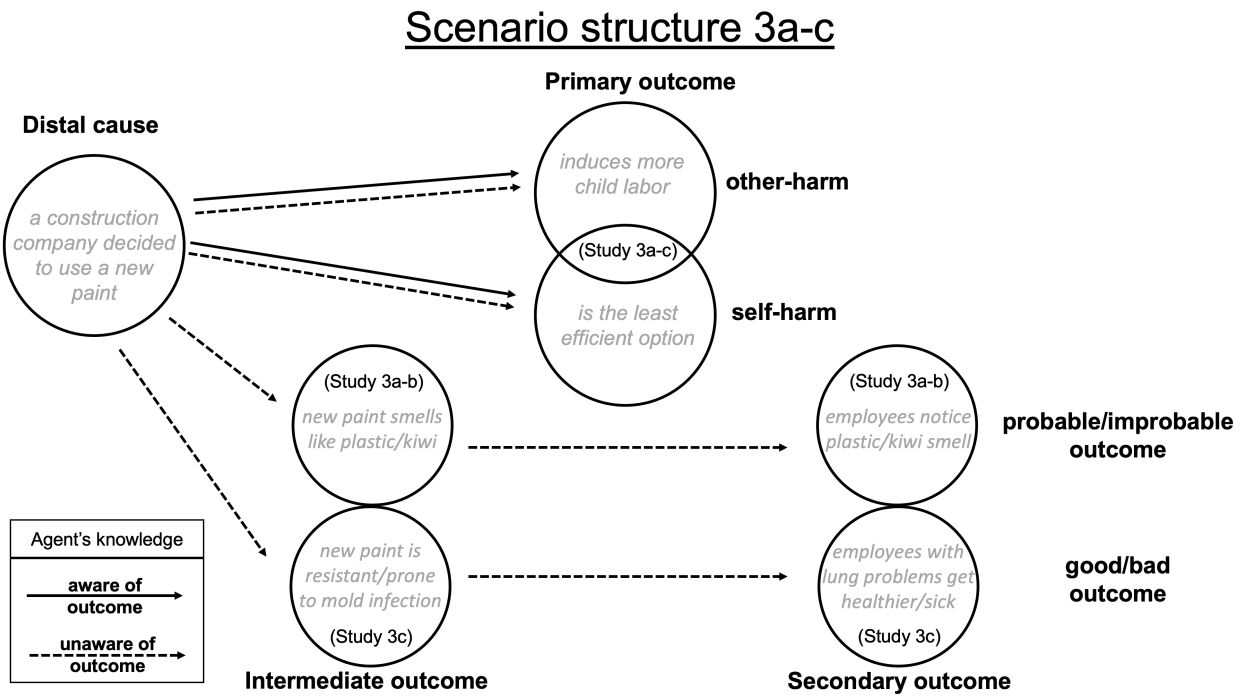


Figure 11: Study design for Experiment series 3

Methods

Participants

For ratings of causation for outcomes that were categorized as probable versus improbable, 227 participants were recruited (M_{age}=34.92, SD_{age}=11.8; 96 females), among which 200 answered all our demographic questions. All participants were recruited through Amazon Mechanical Turk (http://www.mturk.com).

Materials

In this experiment, participants completed 16 trials which each involved reading a vignette where an agent's action eventually resulted in an outcome that was probable or improbable. Separately, the action did harm to others or to the agents themselves, about which the agent was either ignorant or knowledgeable. It is important to point out that knowingly causing harm to others is a violation of moral norms, and knowingly acting in disadvantage to oneself is a rational norm violation. The study used a 2 (Harm Type) × 2 (Agent Knowledge) × 2 (Outcome Probability) × 16 (Scenario) design that was administered in a mixed within- and between-subjects fashion, such that participants saw all 16 scenarios and on each trial were randomly assigned to read 1 of the 8 different versions of that scenario.

For instance, participants may read a vignette as follows:

> **Scenario 1 / Rational Norm / Knowledgeable Agent / Abnormal Outcome** : The owner of a construction company, who was hired to remodel a skyscraper, decided to use a new material called Plyvex to cover the inner walls of the building. When the building inspector reviewed the plans, she realized that Plyvex was one of the least efficient choices for this project. She told the owner of the construction company about this. Although the owner of the construction company understood this, he decided to use Plyvex anyway, like he originally planned. It turns out that Plyvex creates a scent that smells almost exactly like ripe kiwis. Because of this, a number of

people in the company noticed that the entire building smelled of kiwis after the Plyvex had been installed.

Procedure

After reading the brief vignette, participants rated their agreement with a statement about the agent causing the outcome, as in the following example:

> *Causal Question*: The owner of the construction company caused the entire building to smell of kiwis.

Participants answered the question on a 7-point Likert scale from 1 ('Completely disagree') to 7 ('Completely agree'), with a midpoint of 4 ('In between').

Participants were asked to complete a brief demographic questionnaire after completing all 16 trials.

Data analysis

No participant was excluded from the analyses as long as the entire study was completed ($n = 202$). We analyzed data in this study using linear mixed-effects models and investigated all significant effects using estimated marginal means. For the linear mixed-effects models, we included a random intercept for each participant and random slopes for each scenario. More specifically, for each scenario we included a slope for each main effect: *Knowledge*, *Outcome*, and *Harm Type*. Based on estimated marginal means analyses, we report both the estimated marginal mean and the 95% confidence interval.

Results

We analyzed participants' causal judgments, which revealed a main effect of whether the agent had *Knowledge* about bringing about harm, $\chi^2(1) = 13.82$, $p < 0.001$, such that ignorant agents were overall seen as less causal ($M = 3.92$, 95% CI = [3.52, 4.31]) than knowledgeable agents ($M = 4.27$, 95% CI = [3.87, 4.67]), $t(15) = -4.52$, $p < .001$. Importantly, we did not observe either a main effect of the type of *Harm* that the agent incurred, $\chi^2(1) = 0.24$, $p = 0.624$, or a main effect of the probability of the *Outcome*, $\chi^2(1) = 0.99$, $p = 0.321$.

We also probed interaction effects and did not find any interaction effects, including the three-way *Knowledge* $\times$ *Harm Type* $\times$ *Outcome* interaction, $\chi^2(1) = 0.04$, $p = 0.837$, and the two-way interactions of *Knowledge* $\times$ *Outcome*, $\chi^2(1) = 0.1$, $p = 0.746$, *Knowledge* $\times$ *Harm Type*, $\chi^2(1) = 0.37$, $p = 0.540$, and *Outcome* $\times$ *Harm Type*, $\chi^2(1) = 1.81$, $p = 0.179$.

Discussion

The results of Experiment 3a provide two key insights. First, we successfully replicated the knowledge effect observed in Experiment 1 and Experiments 2a–2c for a new set of downstream consequences. Specifically, agents who knowingly engaged in actions that brought harm to themselves or others were considered more causal for the unrelated and unforeseen secondary outcome. Second, we found the anticipated interaction effect between knowledge and the probability of the secondary outcome: knowledge was more influential on causal judgments when the downstream secondary consequence was improbable than when it was probable. This interaction, in we observed in cases that vary descriptive normality, provides further evidence for a unified normality-account for causal judgments.

While this study provides good initial evidence for a more general interaction between outcome normality and agent knowledge, it remains possible that participants still inferred that that the improbable outcomes had

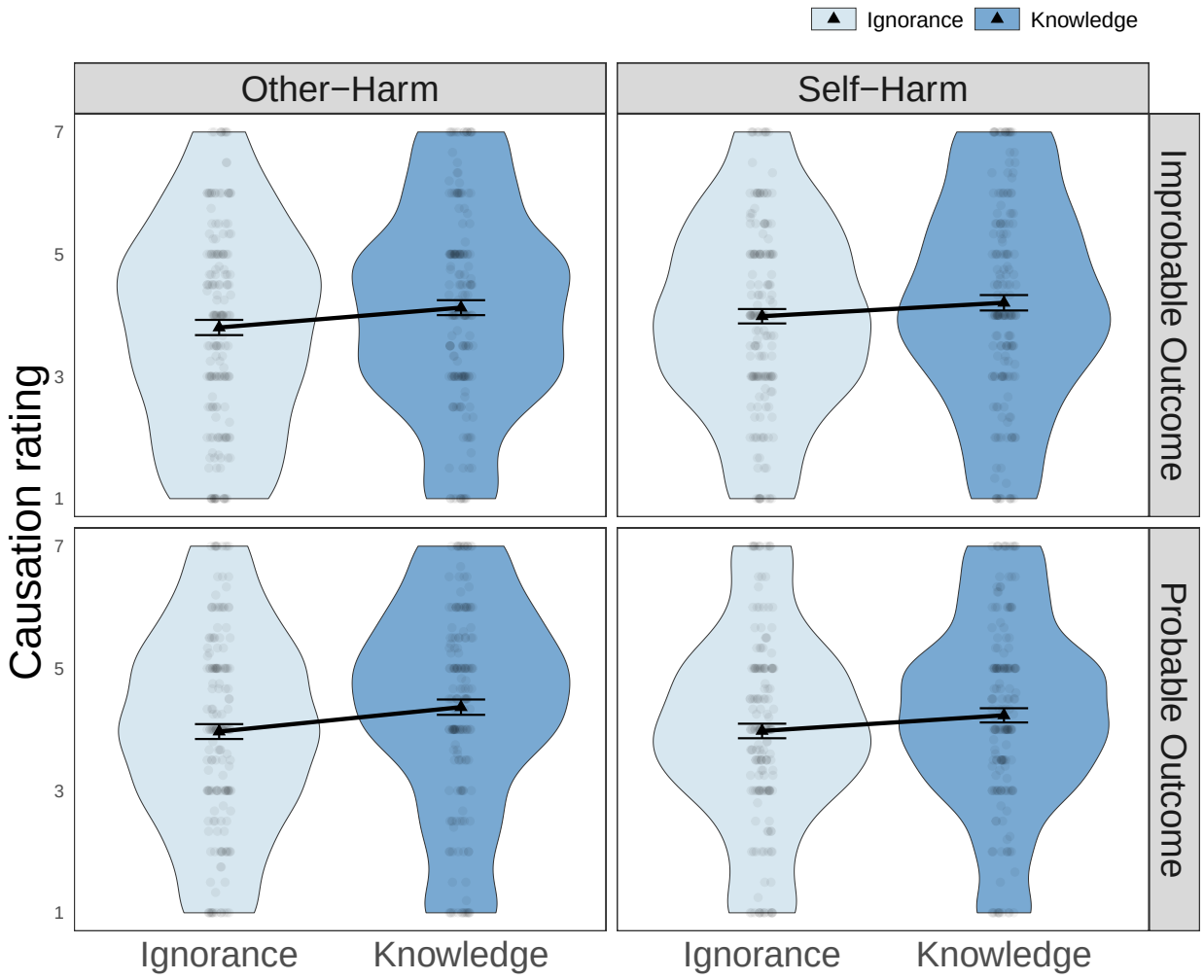


Figure 12: Participants' ratings of causation regarding knowledgeable/ignorant agents who violated moral/rational norms when a separate outcome occurred that is likely/unlikely.

a more negative valence than the probable outcomes. To more precisely investigate whether the interaction effect was driven by outcome valence or outcome normality, we next asked participants to make valence and normality judgments for probable and improbable outcomes in order to directly gauge how their perceptions interact in the context of descriptive normality.

Experiment 3b: Valence and normality judgments in cases of probable/improbable outcomes and norm violations

Using the same causal chains and scenarios as Experiment 3a, we asked participants to provide valence and normality judgments in Experiment 3b, specifically, to what extent they thought the outcome was a good thing and to what extent they thought it was a normal thing to have happened. Doing so serves two purposes. First, we can use these ratings as a manipulation check on the conditions and scenarios we constructed and tested in Experiment 3a. Second, we can use these ratings to capitalize on any variance in perceived normality and valence in the scenarios we constructed. This variance will allow us to more parametrically investigate whether the feature of the outcome that interacts with the agent's knowledge is the valence or the normality of the outcome.

Methods

Participants

For ratings of normality and valence in final outcomes that were probable or improbable, 174 participants were recruited (M_{age}=33.27, SD_{age}=10.61; 63 females), among which 153 answered all our demographic questions. All participants were recruited through Amazon Mechanical Turk (http://www.mturk.com).

Materials

In this experiment, participants completed 16 trials which each involved reading a vignette where an agent's action resulted in a separate secondary outcome that was probable or improbable. The action would have brought about harm to the agent or other people. In addition, the agent was either ignorant or knowledgeable about the harm done. As in Study 3a, the study used a 2 (Harm Type) $\times$ 2 (Agent Knowledge) $\times$ 2 (Outcome Probability) $\times$ 16 (Scenario) design that was administered in a mixed within- and between-subjects fashion, such that participants saw all 16 scenarios and on each trial were randomly assigned to read 1 of the 8 different versions of that scenario.

For instance, participants may read a vignette as follows:

> **Scenario 1 / Rational Norm / Knowledgeable Agent / Improbable Outcome** : The owner of a construction company, who was hired to remodel a skyscraper, decided to use a new material called Plyvex to cover the inner walls of the building. When the building inspector reviewed the plans, she realized that Plyvex was one of the least efficient choices for this project. She told the owner of the construction company about this. Although the owner of the construction company understood this, he decided to use Plyvex anyway, like he originally planned. It turns out that Plyvex creates a scent that smells almost exactly like ripe kiwis. Because of this, a number of people in the company noticed that the entire building smelled of kiwis after the Plyvex had been installed.

Procedure

After reading the brief vignette, participants rated the extent to which they thought the outcome was good or bad and the extent to which they thought the outcome was normal or abnormal.

> *Valence Question*: After the plyvex was installed, the entire building smelled of kiwis. Do you think this was a good or bad thing to have happened?

> *Normality Question*: After the plyvex was installed, the entire building smelled of kiwis. Do you think it was a normal or abnormal thing to have happened?

Participants answered the question on a 7-point Likert scale from 1 ('Completely bad' or 'Completely abnormal') to 7 ('Completely good' or 'Completely normal'), with a midpoint of 4 ('In between').

Participants were asked to complete a brief demographic questionnaire after completing all 16 trials.

Data analysis

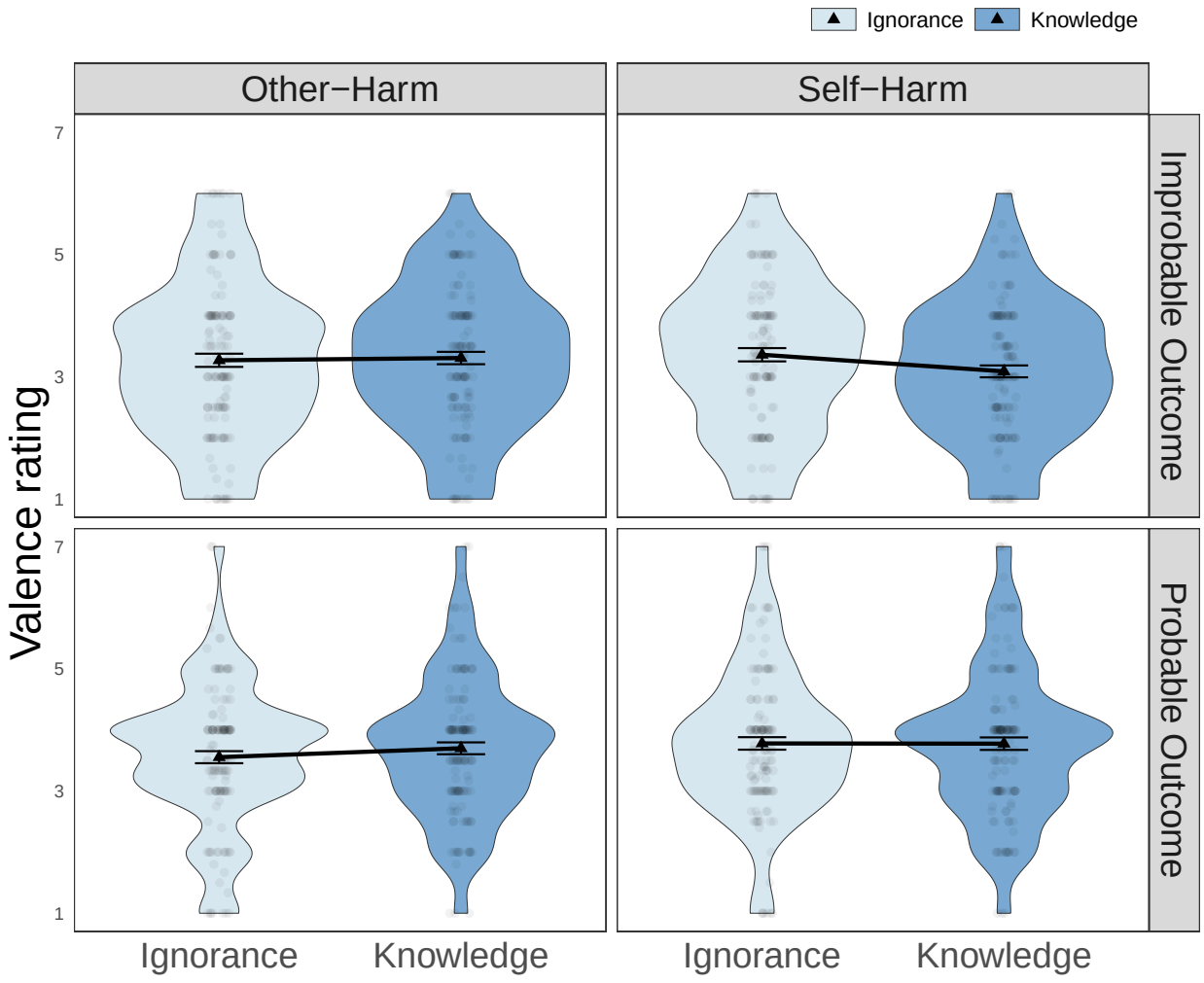
No participant was excluded from the analyses as long as the entire study was completed ($n = 155$). The primary analyses were conducted with linear mixed-effects models and estimated marginal means. More specifically, we included a random intercept for each participant and a random slope for the interaction between *Knowledge* and *Outcome* (or *Knowledge* and *Outcome* specified separately for model convergence) nested in each scenario. We did not include *Harm Type* in the random slope due to its minimal effect. Based on estimated marginal means analyses, we report both the estimated marginal means and the 95% confidence intervals.

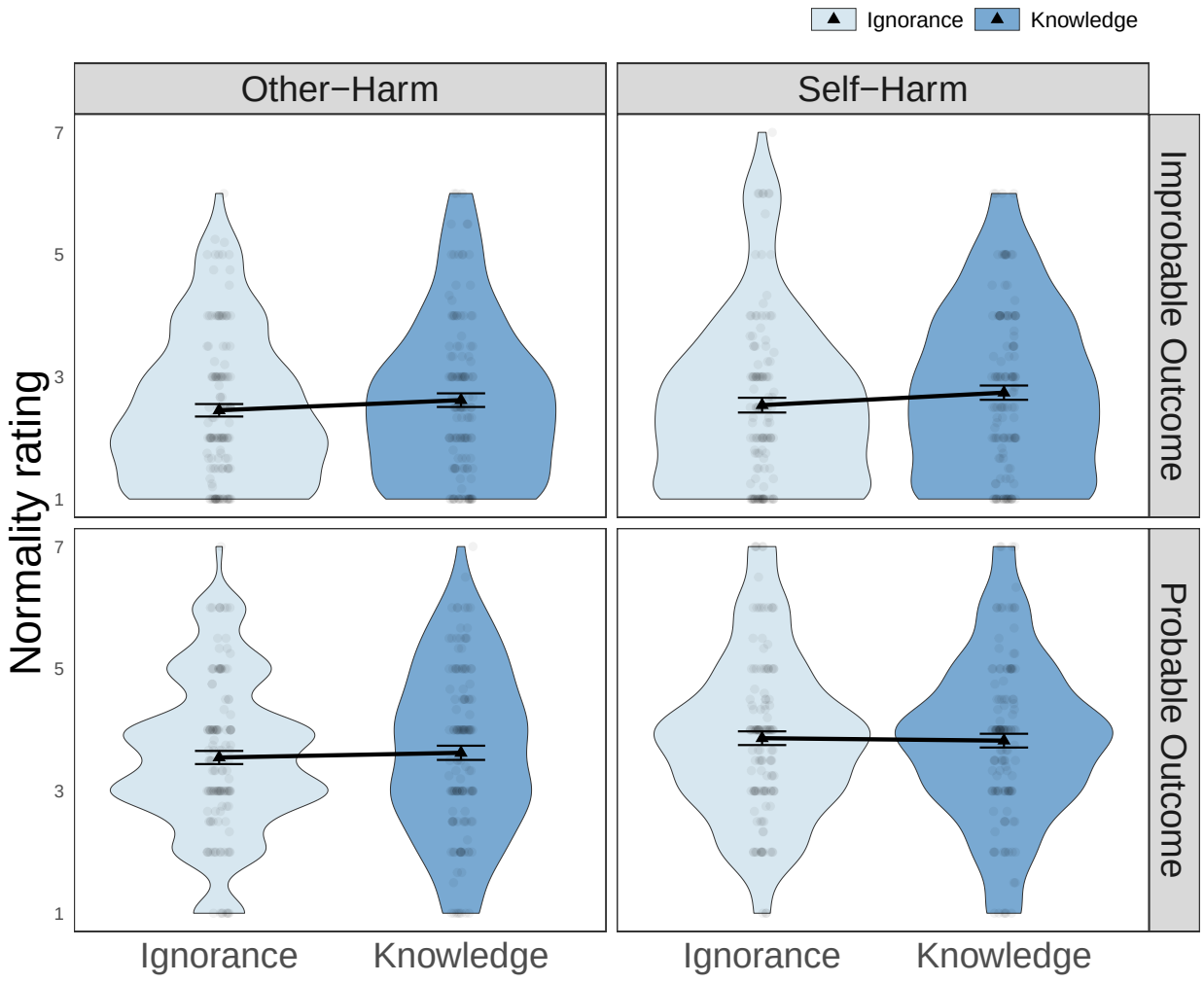
Results

Note that we report all analyses conducted for this experiment to ensure comprehensive coverage, including results that are not directly relevant to our primary predictions.

We analyzed participants' valence ratings, which revealed a main effect of the type of *Harm Type*, $\chi^2(1) = 3.94$, $p = 0.047$, such that participants thought it was less good of a thing for the secondary outcome to have happened when the separate primary outcome caused harm to others ($M = 3.43$, 95% CI = [3.09, 3.76]) than when harm was inflicted on the agents themselves ($M = 3.53$, 95% CI = [3.19, 3.86]), $t(2351.08) = -1.99$, $p = 0.047$. We did not observe either a main effect of whether or not the agent had *Knowledge* about the outcome, $\chi^2(1) = 0.63$, $p = 0.429$, or a main effect of the type of the *Outcome*, $\chi^2(1) = 2.31$, $p = 0.128$. We also probed interaction effects and two stood out. The interaction of *Knowledge* × *Harm Type*, $\chi^2(1) = 5.73$, $p = 0.017$, revealed that when the harmful primary consequence was related to others, more positive valence was attributed to knowledgeable agents ($M = 3.47$, 95% CI = [3.12, 3.81]) than the oblivious ones ($M = 3.39$, 95% CI = [3.04, 3.73]), $t(59.39) = -1.10$, $p = 0.689$; when the harm was related to the agents themselves, occurrences involving knowledgeable agents ($M = 3.44$, 95% CI = [3.10, 3.79]) were considered less positive than those involving ignorant agents ($M = 3.61$, 95% CI = [3.27, 3.95]), $t(60.54) = 2.27$, $p = 0.117$. *Knowledge* × *Outcome* interaction was marginally significant, $\chi^2(1) = 3.64$, $p = 0.057$, suggesting that when secondary outcomes were probable, there was not a large difference in valence attribution between ignorant agents ($M = 3.63$, 95% CI = [3.20, 4.05]) and knowledgeable agents ($M = 3.68$, 95% CI = [3.26, 4.10]), $t(59.61) = -0.73$, $p = 0.886$, whereas in improbable secondary outcomes, those caused by ignorant agents ($M = 3.37$, 95% CI = [2.90, 3.73]) were perceived as better than those caused by knowledgeable agents ($M = 3.23$, 95% CI = [2.75, 3.81]), $t(58.45) = 1.91$, $p = 0.234$. We did not, however, see statistically significant effect in the three-way *Knowledge* × *Harm Type* × *Outcome* interaction, $\chi^2(1) = 1.79$, $p = 0.181$ or the *Outcome* × *Harm Type* interaction, $\chi^2(1) = 0.42$, $p = 0.515$.

We then looked into participants' normality ratings and found main effects of both *Harm Type* and *Outcome*. For *Harm Type*, $\chi^2(1) = 8.88$, $p = 0.003$, participants believed the secondary outcome was less normal when the acting agent brought about harm to others ($M = 3.07$, 95% CI = [2.83, 3.31]) than to themselves ($M = 3.25$, 95% CI = [3.01, 3.49]), $t(2350.47) = -3.00$, $p = 0.003$. Unsurprisingly for *Outcome*, $\chi^2(1) = 40.04$, $p < .001$, participants saw events that were unlikely to happen as less normal ($M = 2.57$, 95% CI = [2.34, 2.79]) than those that were likely to happen ($M = 2.57$, 95% CI = [2.34, 2.79]), $t(14.98) = -11.65$, $p < .001$. There was no main effect concerning *Knowledge* for normality ratings $\chi^2(1) = 1.86$, $p = 0.173$. Analyses did not reveal any statistically significant interaction effects, including the three-way *Knowledge* × *Harm Type* × *Outcome* interaction, $\chi^2(1) = 0.57$, $p = 0.449$, the *Harm Type* × *Outcome* interaction, $\chi^2(1) = 0.95$, $p = 0.329$, the *Knowledge* × *Harm Type* interaction, $\chi^2(1) = 0.00$, $p = 0.959$, and the *Outcome* × *Knowledge* interaction, $\chi^2(1) = 1.10$, $p = 0.294$.


Ignorance
Knowledge
Other−Harm
Self−Harm
Improbable Outcome
Probable Outcome
Valence rating
7
5
3
1
Ignorance
Knowledge
Ignorance
Knowledge



Discussion

In this experiment, we collected participants' judgments of valence and normality of the secondary outcomes that were either probable or improbable. As expected, improbable outcomes were perceived as less normal than probable outcomes; however, they were not necessarily judged as less positive in valence. This finding suggests that the interaction effect of outcome probability and agent knowledge previously observed in Experiment 3a cannot be simply explained by improbable outcomes being perceived as negative in valence. Moreover, readers may recall that the differences in causal judgments were more pronounced between knowledgeable and ignorant agents for both bad outcomes (as opposed to good outcomes) and improbable outcomes (as opposed to probable outcome). Taken together, these results suggest that valence and probability should not be conflated with one another, as valence ratings (observed here) cannot explain the interaction effect observed for probable vs. improbable outcomes. Instead, a more likely explanation is that both effects should both be understood within a broader framework of *normality*.

In the next experiment we follow this suggestion by asking whether the interaction effect in the case of bad vs. good outcomes may be able to be accounted for by differences in pereceived normality. Specifically, we return to manipulating outcome valence (rather than probability) and again ask participants to evaluate valence and normality for good and bad outcomes. If bad outcomes are indeed seen as less normal, we can infer that outcome valuence is influencing perceptions of outcome normality, and therefore move one step closer to a more parsimonious account of normality being impacted by both statistical probability and valence.

Experiment 3c: Valence and normality judgments in cases of bad/good outcomes and norm violations

Mirroring Experiment 3b, we once again asked participants to provide their ratings of outcome-valence and outcome-normality, but this time returned to the scenarios from Experiments 1-2, where we intended to manipulate the valence of the outcome. While the prior ratings allowed us to demonstrate that outcome-normality can matter on its own, these additional ratings enable us to ask the further question of whether the original interaction effect we observed in Experiment 2 may in fact have arisen because of perceived differences in the normality of the outcome rather than perceived differences in the valence of the outcome.

Methods

Participants

For ratings of normality and valence in final outcomes that were either good or bad, 172 participants were recruited (M_{age}=35.25, SD_{age}=10.39; 67 females), among which 152 answered all our demographic questions. All participants were recruited through Amazon Mechanical Turk (http://www.mturk.com).

Materials

In this experiment, participants completed 16 trials which each involved reading a vignette where an agent's action resulted in an outcome that was good or bad and the action would have harmed someone else or the agent themself. In addition, the agent was either ignorant or knowledgeable about the harm. As in Study 3a and Study 3b, the study used a 2 (Harm Type) × 2 (Agent Knowledge) × 2 (Outcome Valence) × 16 (Scenario) design that was administered in a mixed within- and between-subjects fashion, such that participants saw all 16 scenarios and on each trial were randomly assigned to read 1 of the 8 different versions of that scenario.

For instance, participants may read a vignette as follows:

> **Scenario 1 / Self-Harm / Knowledgeable Agent / Good Outcome** : The owner of a construction company, who was hired to remodel a skyscraper, decided to use a new material called Plyvex to cover the inner walls of the building. When the building inspector reviewed the plans, she realized that Plyvex was one of the least efficient choices for this project. She told the owner of the construction company about this. Although the owner of the construction company understood this, he decided to use Plyvex anyway, like he originally planned. It turns out that Plyvex is particularly resistant to mold. Because of this, a number of people in the company who were suffering from lung problems became much healthier after the Plyvex had been installed.

Procedure

After reading the brief vignette, participants rated the extent to which they thought the outcome was good or bad and the extent to which they thought the outcome was normal or abnormal.

> *Valence Question*: After the plyvex was installed, the entire building smelled of kiwis. Do you think this was a good or bad thing to have happened?

> *Normality Question*: After the plyvex was installed, the entire building smelled of kiwis. Do you think it was a normal or abnormal thing to have happened?

Participants answered the question on a 7-point Likert scale from 1 ('Completely bad' or 'Completely abnormal') to 7 ('Completely good' or 'Completely normal'), with a midpoint of 4 ('In between').

Participants were asked to complete a brief demographic questionnaire after completing all 16 trials.

Data analysis

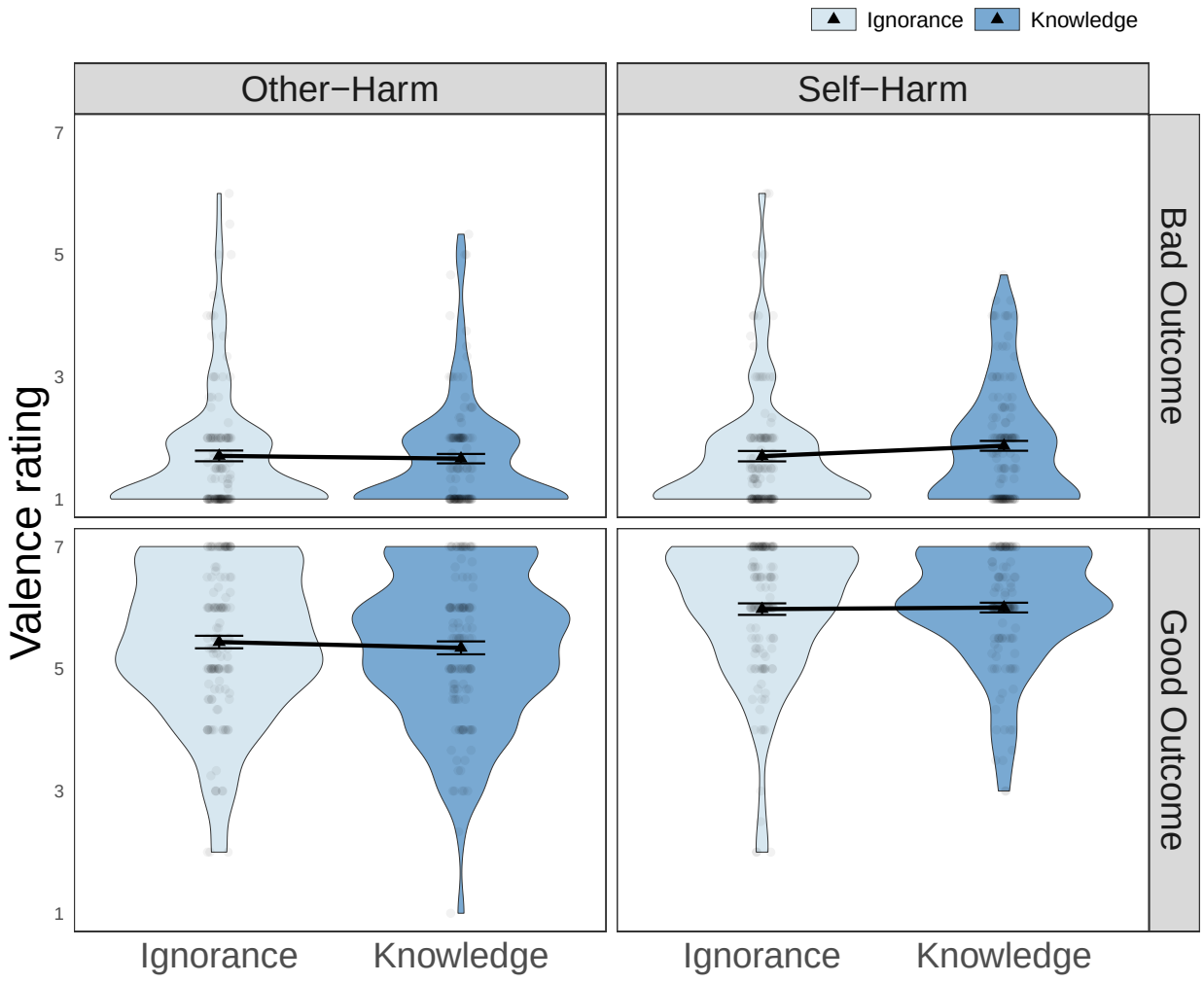
No participant was excluded from the analyses as long as the entire study was completed (n = 153). The primary analyses were conducted with linear mixed-effects models and estimated marginal means. More specifically, we included a random intercept for each participant and a random slope for the interaction between *Knowledge* and *Outcome* nested in each scenario. We did not include *Harm Type* in the random slope due to its minimal effect. Based on estimated marginal means analyses, we report both the estimated marginal means and the 95% confidence intervals.

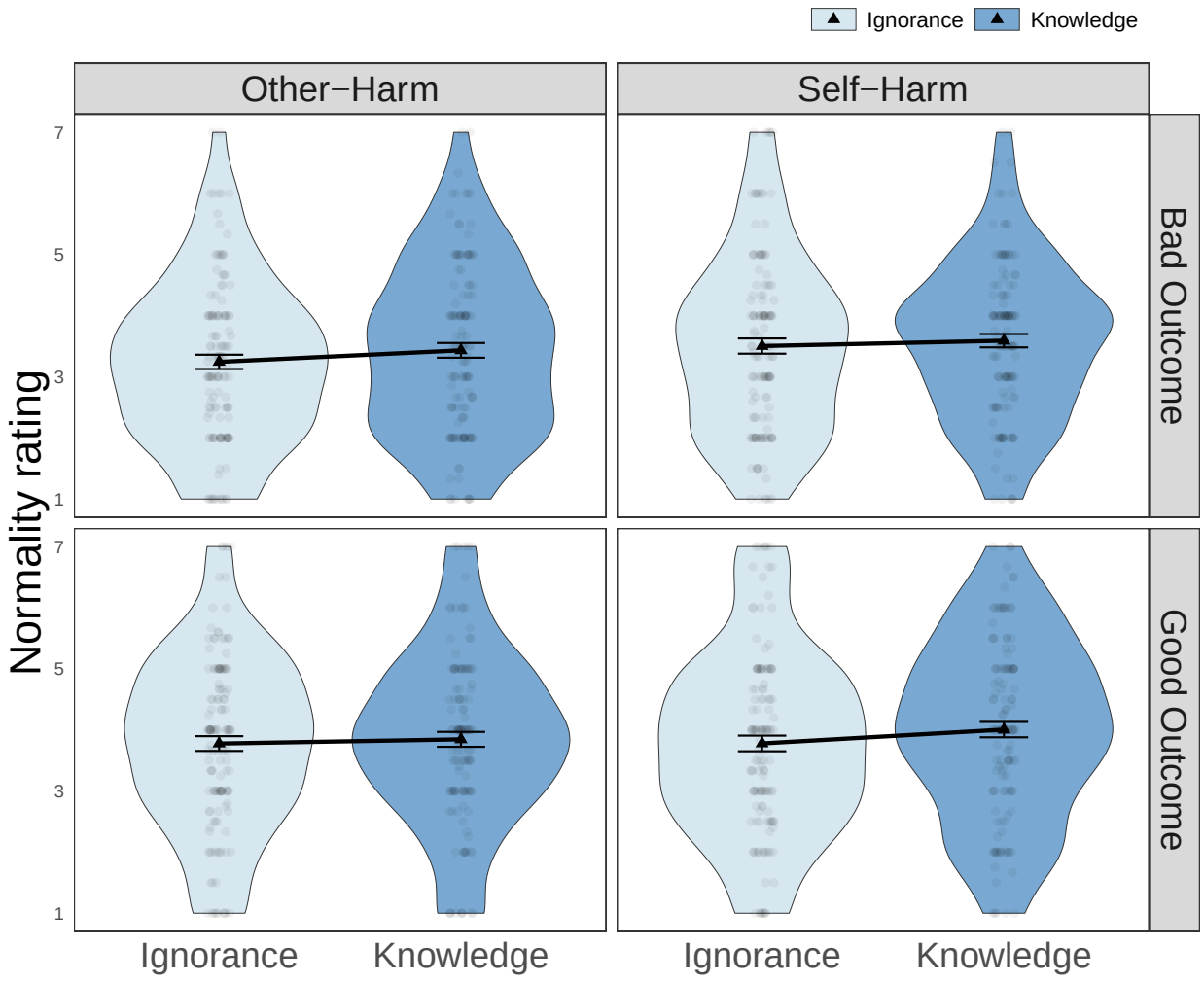
Results

Note that we report all analyses conducted for this experiment to ensure comprehensive coverage, including results that are not directly relevant to our primary predictions.

Analysis of participants' valence ratings revealed a main effect of *Harm Type*, $\chi^2(1) = 51.92$, $p < .001$, where harm to others were seen as less good (M = 3.53, 95% CI = [3.38, 3.68]) than harm to selves (M = 3.88, 95% CI = [3.73, 4.03]), $t(2345.14) = -7.30$, $p < .001$. And as expected, we found a main effect of *Outcome*, $\chi^2(1) = 70.09$, $p < .001$, such that participants agreed it was less good of a thing to have happened when the final outcome was bad (M = 1.73, 95% CI = [1.57, 1.90]) than when it was good (M = 5.67, 95% CI = [5.46, 5.88]), $t(14.96) = -34.42$, $p < .001$. Moreover, there exists an interaction between *Harm Type* and *Outcome*, $\chi^2(1) = 30.23$, $p < .001$, showing that in cases where final the outcome was bad, participants tended to think that harming others was less acceptable (M = 1.69, 95% CI = [1.52, 1.87]) than harming oneself (M = 1.78, 95% CI = [1.6, 1.95]), $t(2346.68) = -1.27$, $p = 0.58$. When the outcome turned out to be good, this difference in valence judgment was even more salient between self-harm (M = 5.98, 95% CI = [5.77, 6.20]) and other-harm (M = 5.36, 95% CI = [5.14, $t(2341.67) = -9.06$, $p < .001$. In addition, the interaction between *Harm Type* and *Knowledge*, $\chi^2(1) = 6.10$, $p = 0.013$, reveals that when acting with knowledge, harming others was considered worse (M = 3.48, 95% CI = [3.31, 3.65]) than harming selves (M = 3.95, 95% CI = [3.78, 4.12], $t(2343.78) = -6.88$, $p < .001$. However, when agents acted obliviously, we saw less of a difference between participants' valence rating of other-harm (M = 3.57, 95% CI = [3.42, 3.73]) and that of self-harm (M = 3.81, 95% CI = [3.65, 3.97]), $t(2341.51) = -3.44$, $p = 0.003$. Testing the main effect of *Knowledge*, $\chi^2(1) = 0.44$, $p = 0.508$, the interaction of *Outcome* × *Knowledge*, $\chi^2(1) = 0.26$, $p = 0.607$, and the three-way interaction of *Knowledge* × *Harm Type* × *Outcome*, $\chi^2(1) = 0.10$, $p = 0.753$, did not yield statistically significant findings.

When we looked into normality rating where we found evidence for all three main effects. The marginally significant *Harm Type* effect, $\chi^2(1) = 3.45$, $p = 0.063$, suggests that participants considered harming others less normal (M = 3.59, 95% CI = [3.33, 3.84]) than harming selves (M = 3.70, 95% CI = [3.45, 3.95]), $t(2308.55) = -1.88$, $p = 0.061$. *Outcome* effect, $\chi^2(1) = 8.62$, $p = 0.003$, indicates that good outcomes were seen as more normal (M = 3.87, 95% CI = [3.55, 4.20]) than bad outcomes (M = 3.41, 95% CI = [3, 3.68]), $t(14.95) = -2.92$, $p = 0.011$. *Knowledge* effect, $\chi^2(1) = 4.10$, $p = 0.043$, reveals that participants thought it was more normal of a thing to have happened when the acting agent was knowledgeable (M = 3.71, 95% CI = [3.45, 3.97]) than when the agent was ignorant (M = 3.58, 95% CI = [3.33, 3.83]), $t(14.93) = -2.04$, $p = 0.060$. It is worth noting that we did not find any interaction effects, including the three-way interaction of *Knowledge* × *Harm Type* × *Outcome*, $\chi^2(1) = 0.07$, $p = 0.791$ and the two-way interactions of *Harm Type* × *Outcome*, $\chi^2(1) = 0.33$, $p = 0.568$, *Harm Type* × *Knowledge*, $\chi^2(1) = 1.83$, $p = 0.176$, *Knowledge* × *Outcome*, $\chi^2(1) = 0.17$, $p = 0.682$.


Ignorance
Knowledge
Other−Harm
Self−Harm
Bad Outcome
Good Outcome
Valence rating
7
5
3
1
Ignorance
Knowledge
Ignorance
Knowledge



Discussion

In this experiment, we collected participants' judgments of valence and normality of the secondary outcomes that were either bad or good. At a broad level, the results from this experiment verified that participants perceived the bad secondary outcomes as more negatively valenced than the good secondary outcomes. Additionally, we also observed that the bad outcomes were perceived as less normal than the good outcomes. This finding supports the suggestion that perceptions of normality are impacted by both descriptive information (the probability of the outcomes) and prescriptive information (the valence of the outcomes). In short, we found that, in support of the more parsimonious account, varying outcome valence changes not only the perceived goodness but perceived normality.

By now, we can reasonably conclude that both valence and probability contribute to perceptions of normality, each shaping how "normal" or "abnormal" an outcome appears. This dual influence on normality, in turn, impacts causal judgments. In the next experiment, we will apply this dual-component theory of normality to construct scenarios with outcomes that vary systematically in both valence and probability, enabling us to examine how these factors jointly contribute to causal cognition.

Experiment 4a: Modality and counterfactuality judgments in cases of bad/good outcomes and norm violations

Across various scenarios we have tested by now, we explored whether participants' causal judgments could be influenced by outcome valence or probability as well as the knowledge status of the acting agent. Prior research suggests that causal judgments are shaped by counterfactual reasoning, often formulated in if-then statements (Petrocelli et al., 2011). These constructs involve two components: the possibility of an antecedent (also known as modality), and the probability that a different antecedent could have led to a different outcome.

Crucially, knowledge affects modality: it is less plausible for people to think that an agent who did something knowingly could have acted in a different way than one who acted ignorantly [CITES?]. Moreover, outcome normality interacts with this knowledge effect in counterfactual reasoning, such that abnormal outcomes (improbable or negative) tend to be perceived as more fragile and susceptible to change, meaning that people see such outcomes as more dependent on the antecedent, especially when the agent knew the consequence of action [BUT we didn't see this interaction with probable/improbable outcomes in 3a so maybe I should rewrite this; also CITES?]. In other words, the more abnormal an outcome is, the more people believe it could change if the circumstances were different.

In Experiment series 4, we shift our focus from causal selection to empirical modal and counterfactual judgments. Modality perception was captured by asking participants "how likely event A could have not happened," while counterfactual reasoning was framed as "if event A did not happen, event B would have been avoided." In Experiment 4a specifically, we aim to understand participants' belief regarding both components of counterfactual thinking when the secondary outcome of an agent's action was positive or negative. The study used a 2 (Harm Type) $\times$ 2 (Agent Knowledge) $\times$ 2 (Outcome Valence) $\times$ 16 (Scenario) design that was administered in a mixed within- and between-subjects fashion, such that participants saw all 16 scenarios and on each trial were randomly assigned to read 1 of the 8 different versions of that scenario.

Besides, we assessed the integration of modal and counterfactual judgments by examining their multiplicative combination, referred to as "potency." Potency captures the degree to which counterfactual reasoning influences causal judgments, reflecting the combined likelihoods of the counterfactual antecedent and its downstream consequence (Petrocelli et al., 2011). We tested this relationship by connecting potency measures with causal judgments from Experiment 2c which used similar scenarios and varied valenced outcomes.

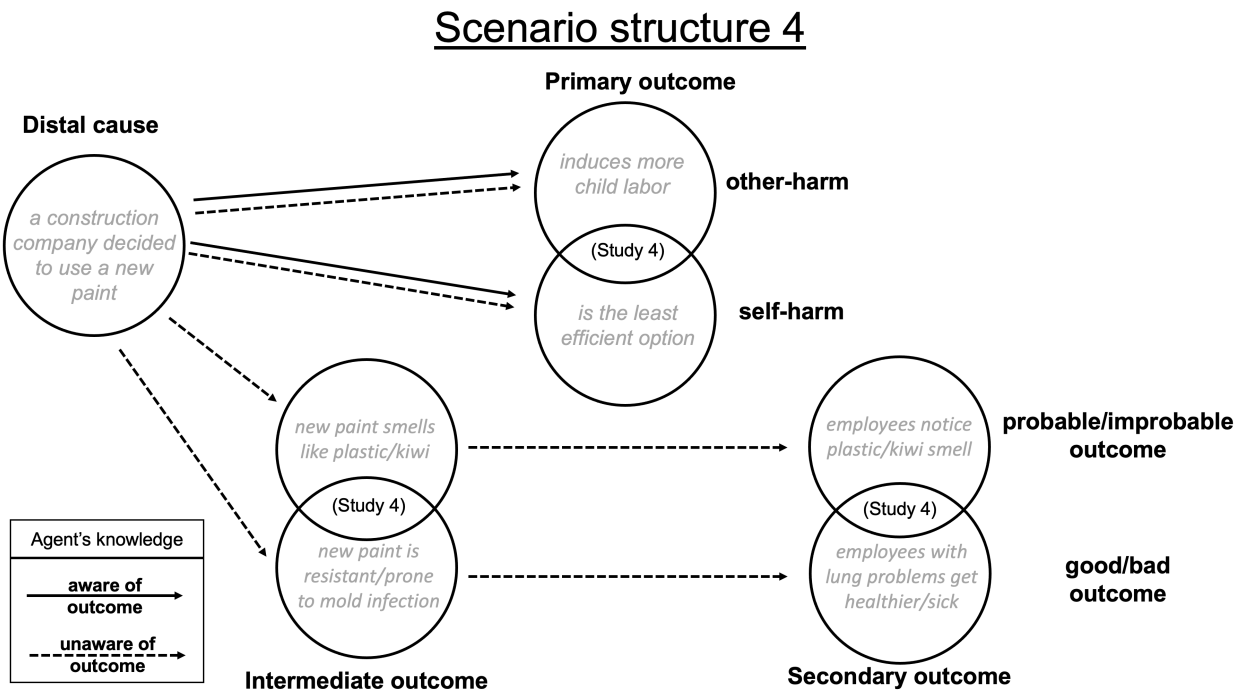


Figure 13: Study design for Experiment 4

Methods

Participants

In Experiment 4a, 224 participants were recruited (M_{age}=34.15, SD_{age}=9.74; 101 females), among which 200 answered all our demographic questions. All participants were recruited through Amazon Mechanical Turk http://www.mturk.com.

Materials

Participants completed 16 trials which each involved reading a vignette where an agent's action resulted in an outcome that was good or bad (valence of occurrence). The action would have harmed either the agent themself or someone else. In addition, the agent was either ignorant or knowledgeable about the norm violated.

For instance, participants may read a vignette as follows:

> **Other-Harm / Knowledgeable Agent / Good Outcome** : The owner of a construction company, who was hired to remodel a skyscraper, decided to use a new material called Plyvex to cover the inner walls of the building. When the building inspector reviewed the plans, she recalled that Plyvex was produced with child slave labor, and purchasing it would cause more children to be enslaved. She told this to the owner. Although the owner of the construction company understood this, he decided to use Plyvex anyway, like he originally planned. It turns out that Plyvex is particularly resistant to mold. Because of this, a number of people in the company who were suffering from lung problems became much healthier after the Plyvex had been installed.

Procedure

After reading each vignette, participants answered two questions: one modal and one counterfactual. More specifically, they rated their agreement with a statement about whether the agent could have refrained from the action, and their belief about the relevance of another statement regarding whether the absence of the agent's action could have prevented the outcome, as in the following example:

> *Modal question*: The owner of the construction company could have not used Plyvex.
>
> *Counterfactual question*: If the owner of the construction company had not used Plyvex, the people wouldn't have recovered.

After completing all 16 trials, participants were asked to complete some optional demographic questions.

Data analysis

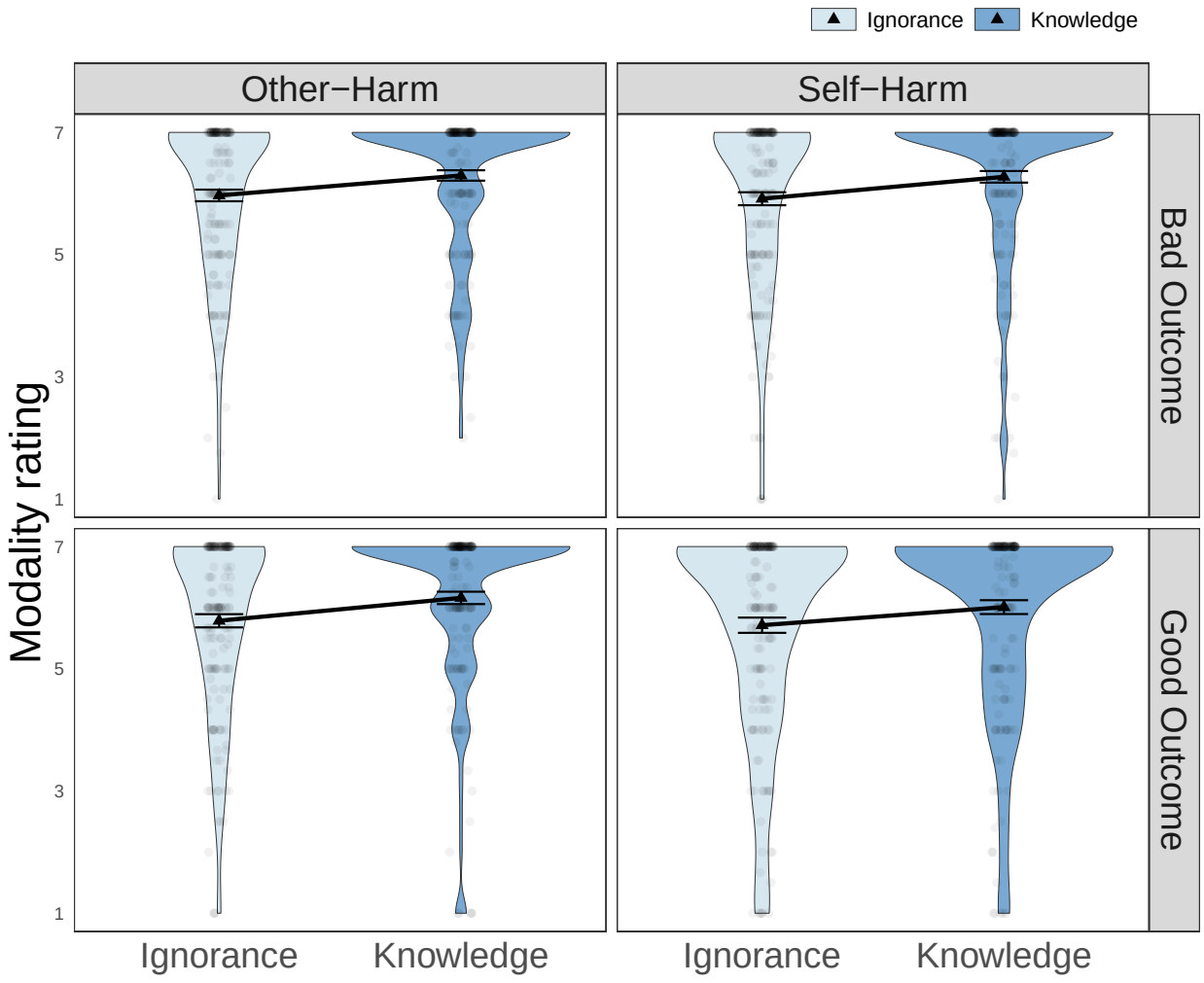
No participants were excluded from the analyses as long as they answered both questions in each of the 16 scenarios. The primary analyses were conducted with linear mixed-effects models and estimated marginal means. More specifically, we included a random intercept for each participant and a random slope for the interaction between *Knowledge* and *Outcome* nested in each scenario. We did not include *Harm Type* in the random slope due to its minimal effect. Based on estimated marginal means analyses, we report both the estimated marginal mean and the 95% confidence interval.

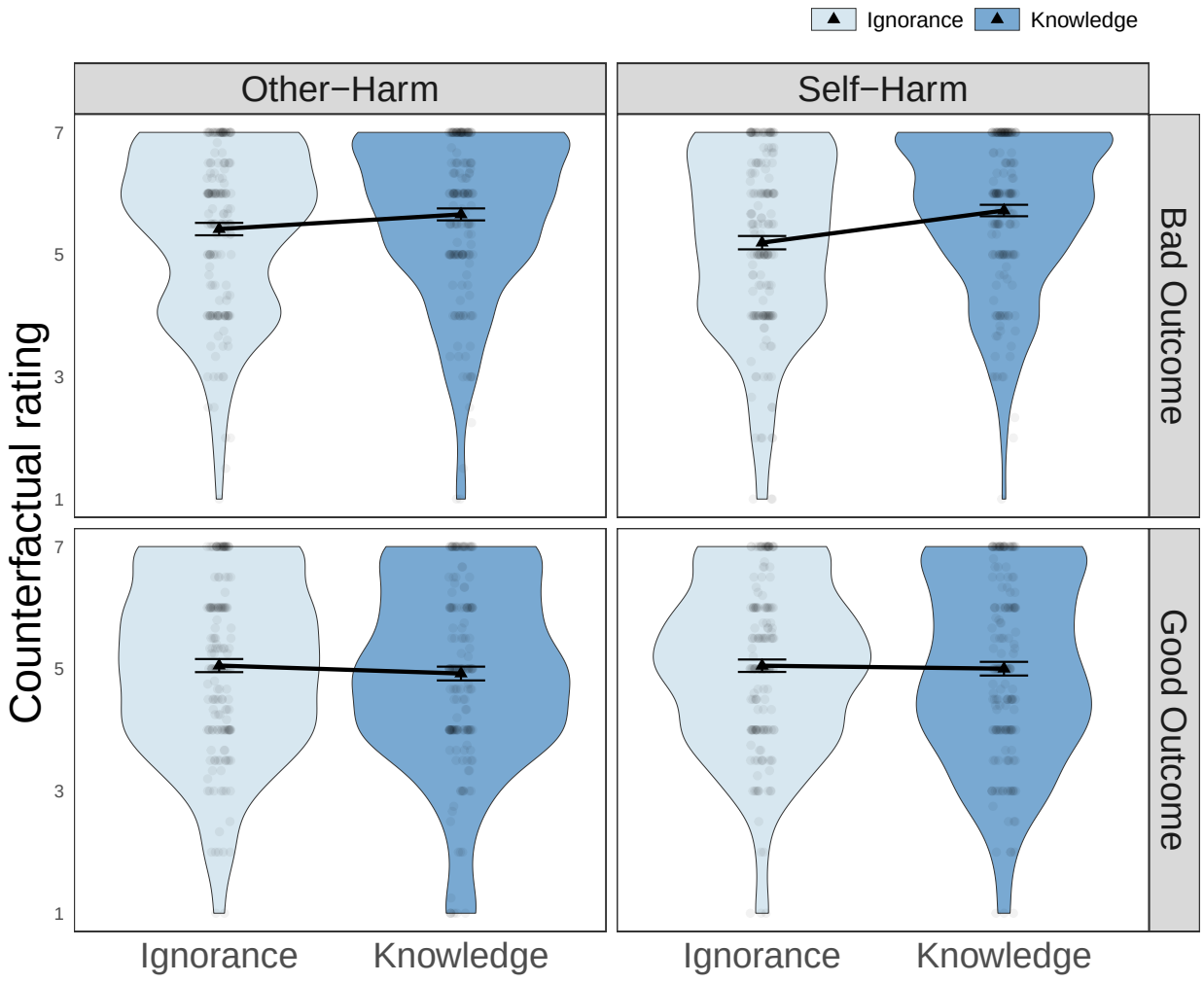
Results

Main effects and interactions effects

We first explored participants' modal ratings and found a significant main effect of *Knowledge*, $\chi^2(1) = 21.85$, $p < .001$, such that participants thought it was less possible for ignorant agents to have not acted in the way they did ($M = 5.86$, 95% CI = [5.67, 6.04]) than knowledgeable agents ($M = 6.19$, 95% CI = [6.03, 6.35]), $t(15.18) = -6.36$, $p < .001$. There was also a significant main effect of *Outcome*, $\chi^2(1) = 8.73$, $p = 0.003$. In other words, participants believed that it was more plausible for the agent to have acted differently when the secondary outcome turned out to be bad ($M = 6.1$, 95% CI = [5.93, 6.28]) than when it turned out to be good ($M = 5.95$, 95% CI = [5.77, 6.12]), $t(15.03) = 3.28$, $p = 0.005$. We did not observe an effect of *Harm Type*, $\chi^2(1) = 1.86$, $p = 0.173$. Analyses of interaction effects in modality ratings did not yield any significant results, including the *Knowledge* × *Harm Type* × *Outcome* interaction, $\chi^2(1) = 1.67$, $p = 0.196$, *Knowledge* × *Outcome* interaction, $\chi^2(1) = 0.07$, $p = 0.798$. We omitted interactions involving the *Harm Type* condition due to its minimal effect.

We then looked into counterfactual ratings, and discovered a significant main effect of *Outcome*, $\chi^2(1) = 16.08$, p NA, $< .001$, such that compared to good outcomes ($M = 4.99$, 95% CI = [4.74, 5.24]), participants were more inclined to think that in bad outcomes, things could have been different if the agents had acted differently ($M = 5.49$, 95% CI = [5.21, 5.77]), $t(15.03) = 6.59$, $p < .001$. Results did not reveal significant main effects in *Knowledge*, $\chi^2(1) = 0.13$, $p = 0.717$ or *Harm Type*, $\chi^2(1) = 0.21$, $p = 0.645$. Analyses did, however, show a significant *Knowledge* × *Outcome* interaction effect in counterfactual ratings, $\chi^2(1) = 6.14$, $p = 0.013$, such that in good secondary outcomes, there was an almost negligible difference in participants' belief that the consequence could have been circumvented with an ignorant agent ($M = 5.03$, 95% CI = [4.75, 5.31]) as opposed to a knowledgeable agent ($M = 4.95$, 95% CI = [4.70, 5.19]), $t(15.01) = 0.96$, $p = 0.775$. Whereas in bad outcomes, participants were more likely to think that when agents were knowledgeable ($M = 5.63$, 95% CI = [5.3, 5.96]), a different action could have resulted in a different outcome than when agents were ignorant ($M = 5.35$, 95% CI = [5.07, 5.63]), $t(15.15) = -2.38$, p 0.123. Again, there was not a significant three-way *Knowledge* × *Harm Type* × *Outcome* interaction effect, $\chi^2(1) = 1.32$, $p = 0.250$, nor any interaction effects involving *Harm Type*.


Ignorance
Knowledge
Other−Harm
Self−Harm
Bad Outcome
Good Outcome
Modality rating
7
5
3
1
Ignorance
Knowledge
Ignorance
Knowledge



Potency and its relationship with causal judgments

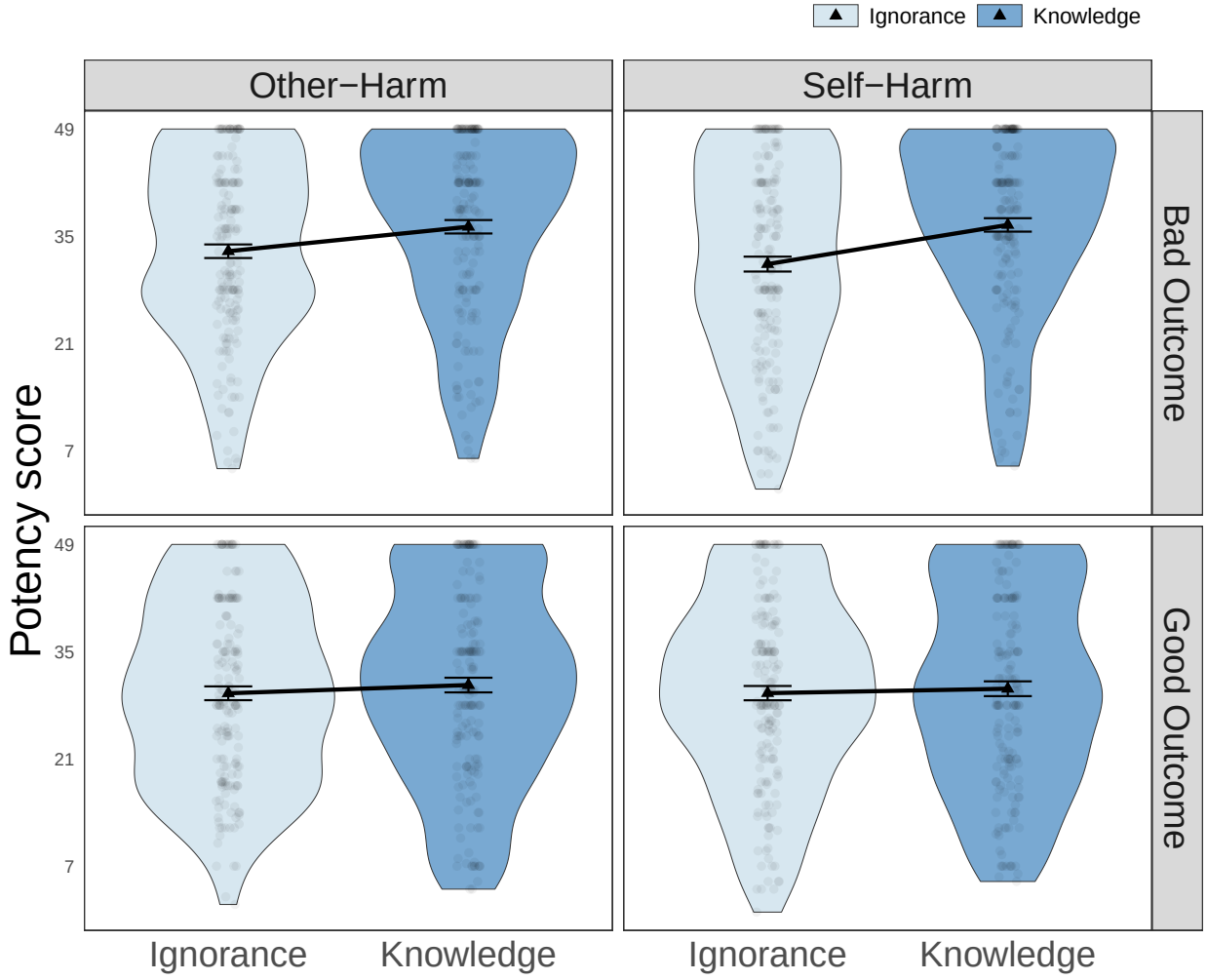
In this part of the analysis, we constructed a measure of potency by calculating the product of participants' modal and counterfactual ratings. This approach allows us to explore the impact of counterfactual reasoning on causal judgments, using data from both the current study and Experiment 2c.

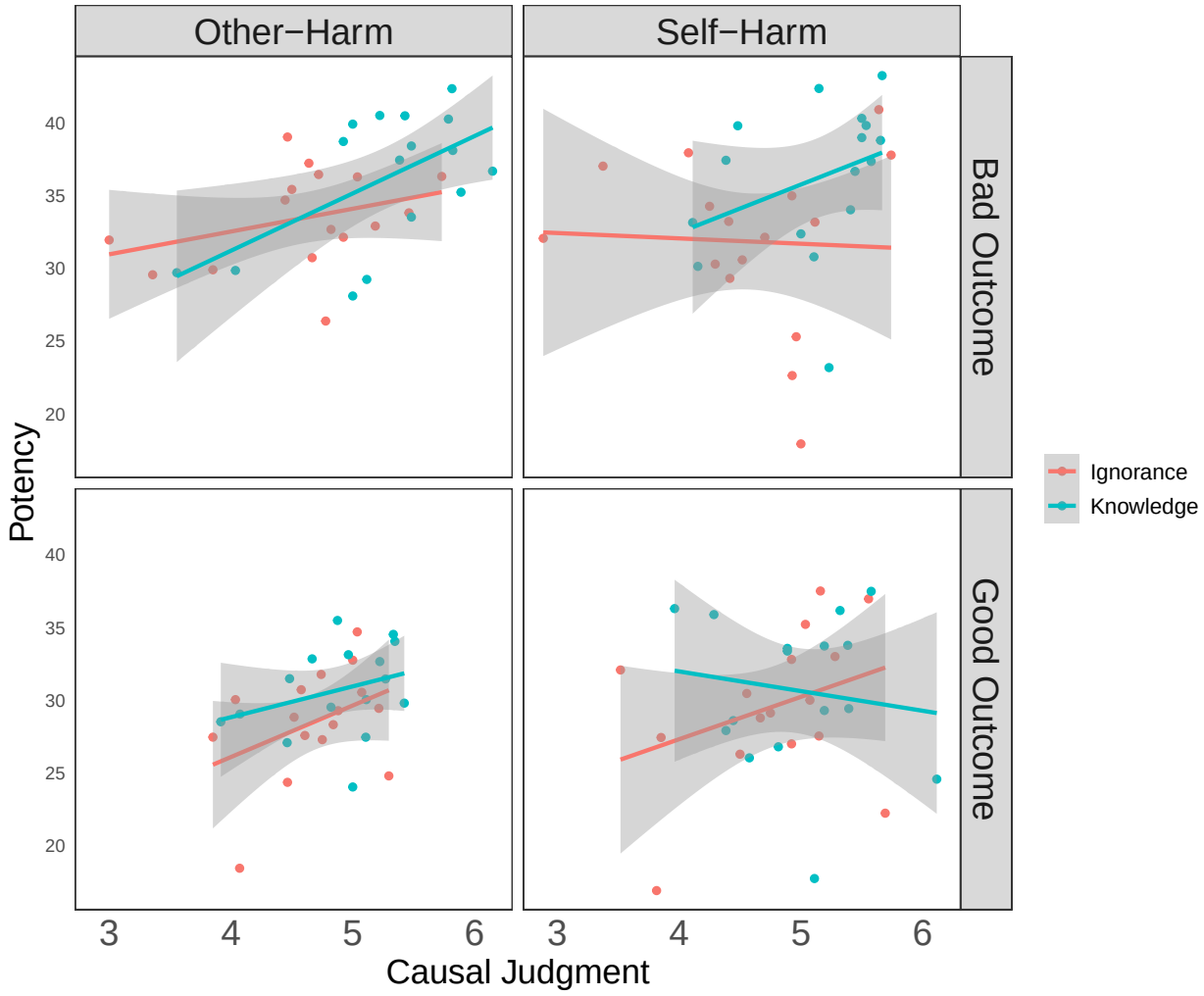
We first looked into main effects of *Knowledge*, *Harm Type*, and *Outcome* on potency and we found statistical significance in both *Knowledge*, $\chi^2(1) = 9.21$, p NA, 0.002, and *Outcome*, $\chi^2(1) = 17.59$, $p < .001$. Potency being higher for knowledgeable agents ($M = 33.24$, 95% CI = [31.19, 35.29]) than ignorant agents ($M = 30.98$, 95% CI = [28.96, 33]) suggests that counterfactual reasoning exerts a larger amount of influence on causal selection on the former type of agents than the latter, $t(15.19) = -4.54$, $p < .001$. Potency was also higher in bad outcomes ($M = 34.19$, 95% CI = [31.95, 36.44]) than in good outcomes ($M = 30.03$, 95% CI = [28.16, 31.90]), $t(15.01) = 6.92$, $p < .001$. In addition, there was a statistically significant *Knowledge* × *Outcome* interaction, $\chi^2(1) = 5.5$, $p = 0.019$, such that in good outcomes, the effect of potency on knowledgeable agents ($M = 30.55$, 95% CI = [28.66, 32.43]) did not differentiate greatly from that on ignorant agents ($M = 29.51$, 95% CI = [27.45, 31.57]), $t(15) = -1.72$, $p = 0.349$. Whereas in bad outcomes, potency had a much larger influence on agents with knowledge ($M = 35.94$, 95% CI = [33.38, 38.5]) versus those without ($M = 32.45$, 95% CI = [30.22, 34.68]), $t(15) = -4.39$, p 0.003.

On the other hand, we did not see significant effects in *Harm Type*, $\chi^2(1) = 1.17$, $p = 0.279$, or the three-way *Knowledge* × *Harm Type* × *Outcome* interaction, $\chi^2(1) = 3.71$, $p = 0.054$. We omitted interactions involving

the *Harm Type* condition due to its minimal effect.

[[[To further corroborate the role of potency measure in causal cognition, we looked into its relationship with causal judgment collected in a similar study (i.e., Experiment 3a). Our analysis shows that potency is a mediator between *Knowledge*, *Outcome Normality* and *Causal Judgment*, ($ACME = .034$, $p < .001$). We also found an interaction between potency and the normality of outcome in predicting causal judgments, $\chi^2(1) = 1.71$, $p = 0.192$.]]]





Discussion

Experiment 4a expands on our understanding of causal cognition in negative and positive secondary outcomes by examining modal and counterfactual judgments, that is, how participants assess the plausibility of alternative antecedent events and the likelihood of different outcomes under such hypothetical alternatives. Modality ratings were influenced by both knowledge and outcome valence, suggesting that the availability of alternatives was driven by whether the agent knowingly caused harm and whether the secondary outcome was negative. Counterfactuals, on the other hand, was impacted by outcome valence and its interaction with knowledge, such that for negative outcomes, the counterfactual reasoning was more dependent on the antecedent when said antecedent involved an agent knowingly causing harm. In consistency with previous studies, these effects revealed that participants were more likely to consider an alternative outcome plausible when the causal agent was knowledgeable rather than ignorant, especially in cases of bad secondary outcomes. Tying this effect back to the same *Knowledge* × *Outcome* interaction previously found studies involving causal judgments prompted us to consider the underlying role of counterfactual reasoning in causal attribution.

We put the two pieces of counterfactual thinking together to understand whether and how they could jointly explain causal judgments, and saw in the potency results that it was impacted by knowledge, outcome valence, and the interaction between the two. This measure integrates perceived likelihood of counterfactual antecedent with the perceived likelihood of counterfactual consequence. The higher it is, the easier it

should be for participants to think of such counterfactuals. Based on the observed effects, we conclude that knowledge status, outcome valuence, and their interaction, all contribute to potency. More specifically, when causal agents acted knowingly as opposed to ignorantly or when secondary outcomes were negative as opposed to positive, counterfactual reasoning had a larger impact on causal judgments, and the said knowledge effect was even more prominent when secondary outcomes were unwelcome. It is also worth noting that potency did not interact with outcome valence or knowledge in predicting causal selection (?: are these still worth including??), suggesting a unidirectional influence coming from outcome valence to potency, eventually to causal judgments. Taken together, our findings support a model where prescriptive normality influences causal judgments through counterfactual reasoning.

Experiment 4b: Modality and counterfactuality judgments in cases of probable/improbable outcomes and norm violations

Following Experiment 4a's exploration of modality and counterfactual judgments with different prescriptive normality, Experiment 4b seeks to extend this investigation into the realm of descriptive normality. The current study used scenarios where secondary outcomes were either likely or unlikely, with a 2 (Harm Type) $\times$ 2 (Agent Knowledge) $\times$ 2 (Outcome Probability) $\times$ 16 (Scenario) design that was administered in a mixed within- and between-subjects fashion, such that participants saw all 16 scenarios and on each trial were randomly assigned to read 1 of the 8 different versions of that scenario.

Once again, we multiplied modality and counterfactual ratings to capture potency, namely the influence of counterfactual reasoning on causal judgments. We then drew causal judgment data from Experiment 3a which used comparable scenarios and varied outcome probability to further evaluate the relationship between the two.

Methods

Participants

In Experiment 4b, 238 participants were recruited (M_{age}=32.38, SD_{age}=10.14; 81 females), among which 200 answered all our demographic questions. All participants were recruited through Amazon Mechanical Turk http://www.mturk.com.

Materials

Participants completed 16 trials which each involved reading a vignette where an agent's action resulted in an outcome that was normal or abnormal (defined by probability of occurrence). The action would have violated a moral or a rational norm. In addition, the agent was either ignorant or knowledgeable about the norm violated.

For instance, participants may read a vignette as follows:

> **Other-Harm / Knowledgeable Agent / Improbable Outcome** : The owner of a construction company, who was hired to remodel a skyscraper, decided to use a new material called Plyvex to cover the inner walls of the building. When the building inspector reviewed the plans, she realized that Plyvex was one of the least efficient choices for this project. She told the owner of the construction company about this. Although the owner of the construction company understood this, he decided to use Plyvex anyway, like he originally planned. It turns out that Plyvex creates a scent that smells almost exactly like ripe kiwis. Because of this, a number of people in the company noticed that the entire building smelled of kiwis after the Plyvex had been installed.

Procedure

After reading each vignette, participants answered two questions: one modal and one counterfactual. More specifically, they rated their agreement with a statement about whether the agent could have refrained from the action, and their belief about the relevance of another statement regarding whether the absence of the agent's action could have prevented the outcome, as in the following example:

> *Modal question*: The owner of the construction company could have not used Plyvex.
>
> *Counterfactual question*: If the owner of the construction company had not used Plyvex, the entire building would not have smelled of kiwis.

After completing all 16 trials, participants were asked to complete some optional demographic questions.

Data analysis

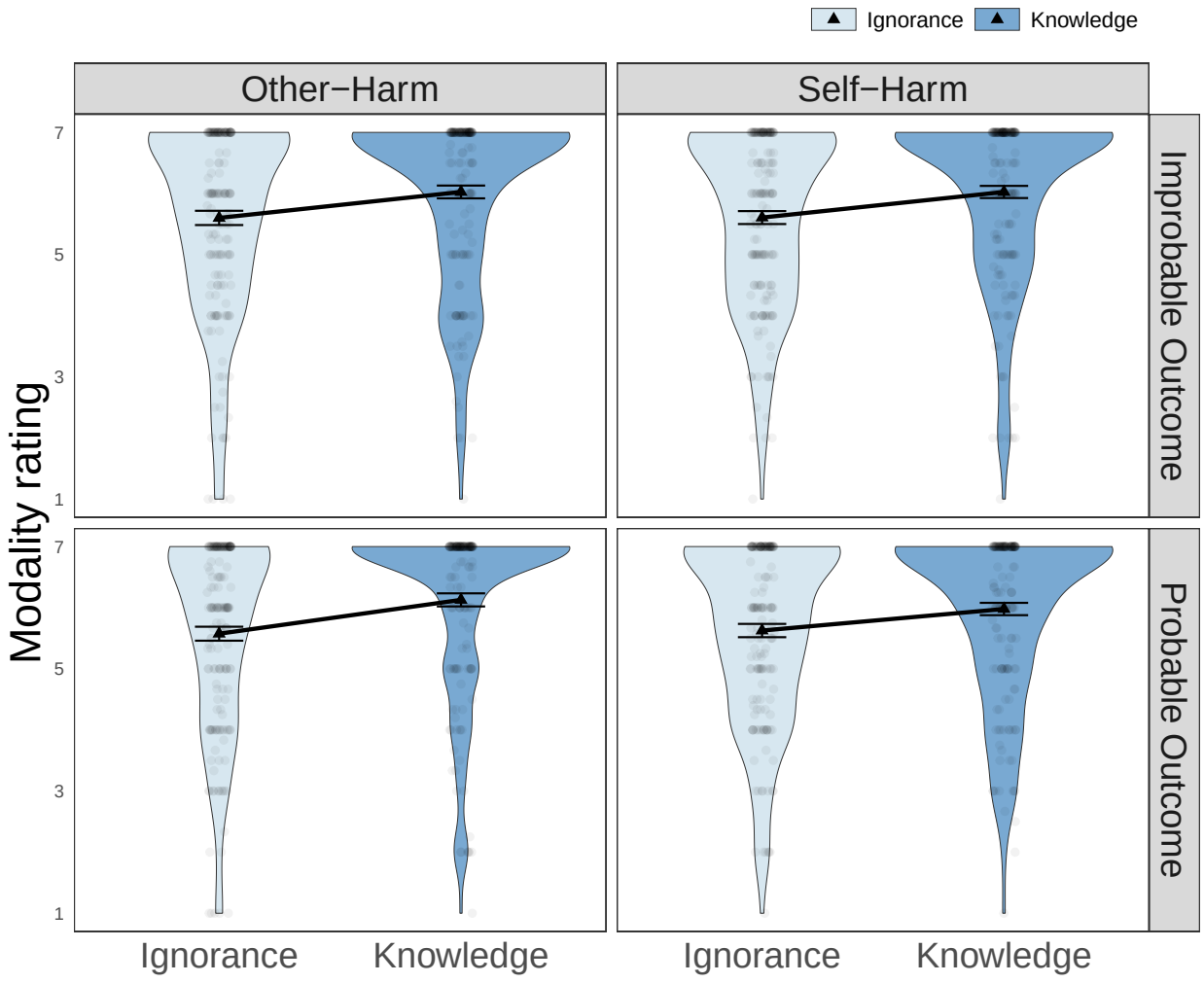
No participants were excluded from the analyses as long as they answered both questions in each of the 16 scenarios. The primary analyses were conducted with linear mixed-effects models and estimated marginal means. More specifically, we included a random intercept for each participant and a random slope for the interaction between *Knowledge* and *Outcome* (or *Knowledge* and *Outcome* as separate covariates for model convergence) nested in each scenario. We did not include *Harm Type* in the random slope due to its minimal effect. Based on estimated marginal means analyses, we report both the estimated marginal means and the 95% confidence intervals.

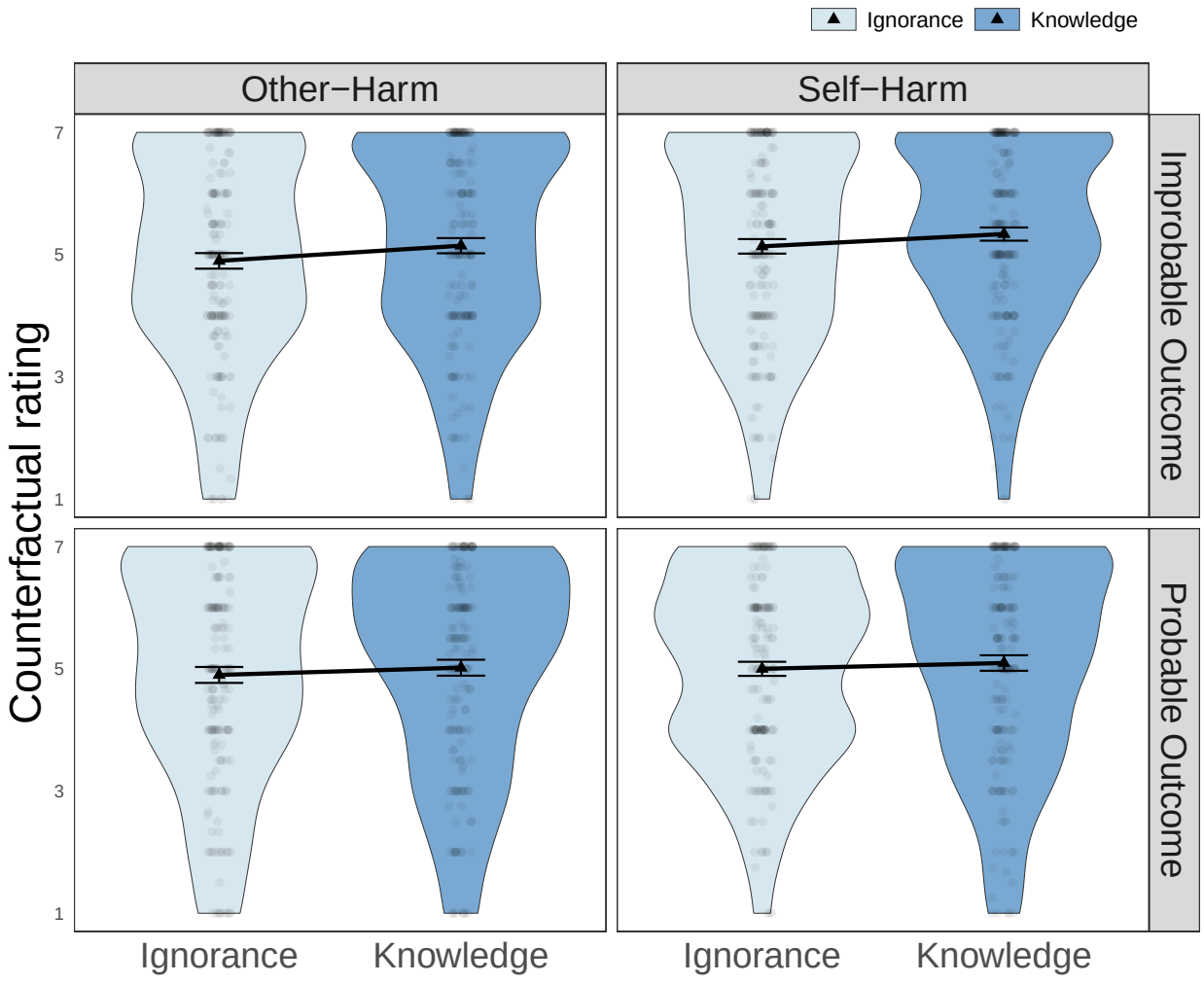
Results

Main effects and interactions effects

We first looked at participants' modal ratings and found a significant main effect of *Knowledge*, $\chi^2(1) = 26.84$, $p < .001$, such that participants thought it was less plausible for ignorant agents to have not acted in the way they did ($M = 5.61$, 95% CI = [5.41, 5.81]) than knowledgeable agents ($M = 6.04$, 95% CI = [5.85, 6.23]), $t(15.19) = -7.82$, $p < .001$. There was no significant main effect in *Outcome*, $\chi^2(1) = 1.35$, $p = 0.245$, or *Harm Type*, $\chi^2(1) = 0.79$, $p = 0.375$. Analyses of interaction effects in modality ratings did not yield any significant results, including the *Knowledge* × *Harm Type* × *Outcome* interaction, $\chi^2(1) = 0.63$, $p = 0.426$, *Knowledge* × *Outcome* interaction, $\chi^2(1) = 0.05$, $p = 0.829$. We omitted interactions involving the *Harm Type* condition due to its minimal effect.

Then, we explored counterfactual ratings, and discovered once again a significant main effect of *Knowledge*, $\chi^2(1) = 5.02$, $p = 0.025$, such that participants were more inclined to believe that things could have been different if a knowledgeable agent had acted differently ($M = 5.03$, 95% CI = [4.75, 5.31]) than when an ignorant agent was involved ($M = 5.15$, 95% CI = [4.86, 5.45]), $t(15.29) = -2.25$, $p = 0.04$. There was also a marginally significant effect of *Harm Type*, $\chi^2(1) = 3.84$, $p = 0.05$: when an action harmed the agent themself, participants considered it more likely that the outcome could have been circumvented if the agent acted differently ($M = 5.15$, 95% CI = [4.86, 5.43]) than when the action brought harm to others ($M = 5.04$, 95% CI = [4.75, 5.32]), $t(3118.45) = -1.96$, $p = 0.05$. There was no significant effect of *Outcome* in counterfactual judgments, $\chi^2(1) = 2.27$, $p = \text{NA}$, $= 0.132$. In addition, we did not find any interaction effects, including the three-way *Knowledge* × *Harm Type* × *Outcome* interaction effect, $\chi^2(1) = 0.24$, $p = 0.626$ and the two-way *Knowledge* × *Outcome* interaction effect, $\chi^2(1) = 0.42$, $p = 0.519$.


Ignorance
Knowledge
Other−Harm
Self−Harm
Improbable Outcome
Probable Outcome
Modality rating
7
5
3
1
Ignorance
Knowledge
Ignorance
Knowledge

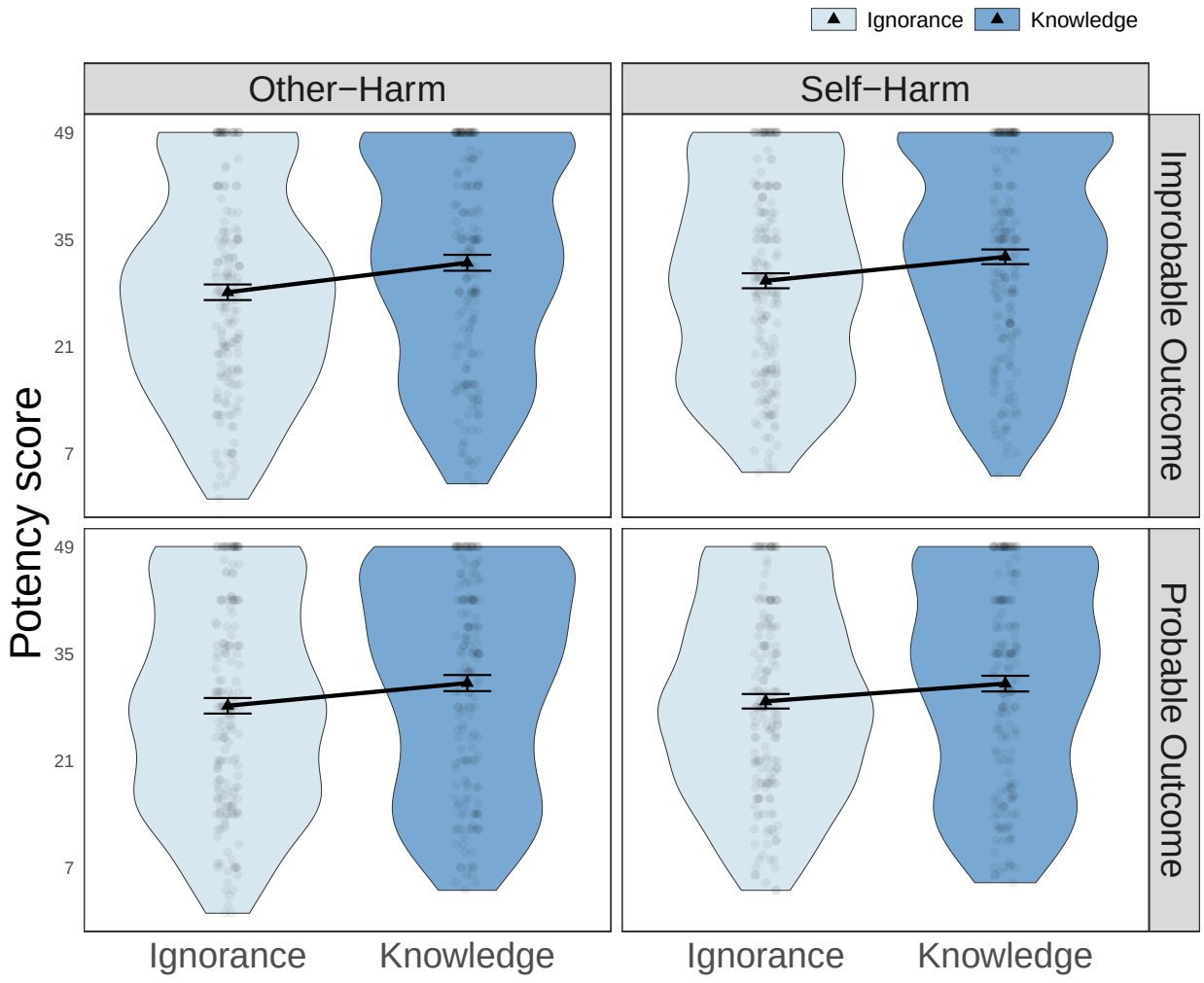


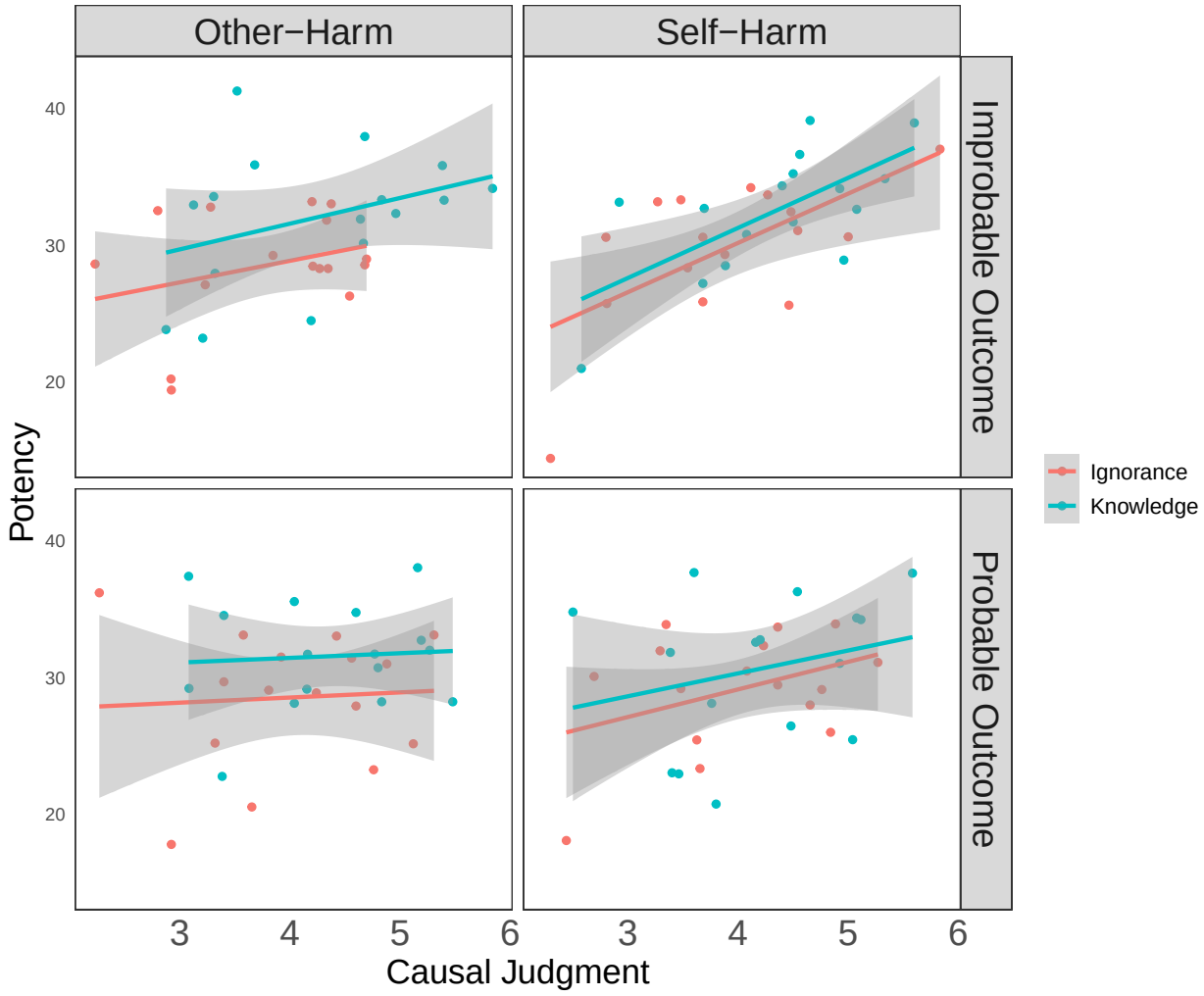
Potency and its relationship with causal judgments

Like in the previous study, we constructed a measure of potency by calculating the product of participants' modal and counterfactual ratings. This approach sheds light on the impact of counterfactual reasoning on causal judgments, using data from both the current study and Experiment 3a.

Examining main effects of *Knowledge*, *Harm Type*, and *Outcome* on potency, we found statistical significance only in *Knowledge*, $\chi^2(1) = 18.14$, p NA, $< .001$, which is to say that counterfactual reasoning has a bigger impact on causal judgment when the agents involved knew the consequence of their actions ($M = 31.89$, 95% CI = [29.67, 34.1]) than when they did not ($M = 29.19$, 95% CI = [26.91, 31.46]), $t(15.21) = -5.74$, $p < .001$. Other than that, we did not see a main effect in either *Outcome*, $\chi^2(1) = 0.51$, p NA, $= 0.476$, or *Harm Type*, $\chi^2(1) = 0.53$, p NA, $= 0.465$. No significant interaction effects were observed, including the three-way *Knowledge* × *Harm Type* × *Outcome* interaction, $\chi^2(1) = 0.12$, $p = 0.730$, and the *Knowledge* × *Outcome* interaction, $\chi^2(1) = 1.25$, $p = 0.264$. We omitted interactions involving the *Harm Type* condition due to its minimal effect.

[[[To further corroborate the role of potency measure in causal cognition, we looked into its relationship with causal judgment collected in a similar study (i.e., Experiment 3a). Our analysis shows that potency is a mediator between *Knowledge*, *Outcome Normality* and *Causal Judgment*, (*ACME* = .034, $p < .001$???). We also found an interaction between potency and the probability of outcome in predicting causal judgments, $\chi^2(1) = 5.36$, $p = 0.021$.]]]


Ignorance
Knowledge
Other−Harm
Self−Harm
Improbable Outcome
Probable Outcome
Potency score
49
35
21
7
Ignorance
Knowledge



Discussion

Experiment 4b examined modality and counterfactual judgments, as well as the magnitude of counterfactual potency, in event chains where the final outcome was either likely or unlikely. Our analyses revealed a consistent *Knowledge* effect through all three measures, indicating that participants reliably associated knowledgeable agents with increased possibility in alternative antecedents and different consequences stemming from said antecedents. Furthermore, when agents acted with knowledge, counterfactual reasoning exerts more influence on assignment of causation. (NOTE: you said no need to test interaction between potency and outcome normality anymore but i tried it out and it's significant in this case, i.e., there's interaction between outcome probability and mean potency in predicting mean causation rating.) The findings resonate with those related to causal judgments in the comparable Experiment 3a where knowledgeable agents were deemed more causal of consequences. Taken together, the results point to an account where knowledge heavily contributes to causal and counterfactual reasoning when actions give rise to descriptively normal or abnormal outcomes.

Appendix

...

Bates, D., Maechler, M., Bolker, B., Walker, S., et al. (2014). lme4: Linear mixed-effects models using eigen and S4. *R Package Version*, *1*.

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

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