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

As seen in the following experiments, this paper made several unique contributions. The first one is the use of causal chains. We constructed events that had clear causal relationships with one another, which is crucial in understanding participants' perception regarding causation and counterfactuals. Second, on top of replicating past work that examined agents who knowingly contributed to moral norm violations, we included ignorant agents and inanimate objects to examine the role of knowledge in causal selection. Furthermore, we no longer confine ourselves to studying moral norm violations; instead, we systematically manipulated valence of outcomes to understand causal judgments under varying circumstances.

In Experiment 1, we used 24 scenarios where agents with different knowledge status contributed to different outcomes. More specifically, all vignettes involved a distal cause, an immediate outcome that is at the same time a proximal cause, and a final outcome. We varied the type of cause (knowledgeable human agent, ignorant human agent, or inanimate object) and the valence of the final outcome (good or bad). We then asked participants to 1) rate how causal the distal cause was to the outcome, 2) between distal and proximal cause, choose the one that could have resulted in a different outcome.

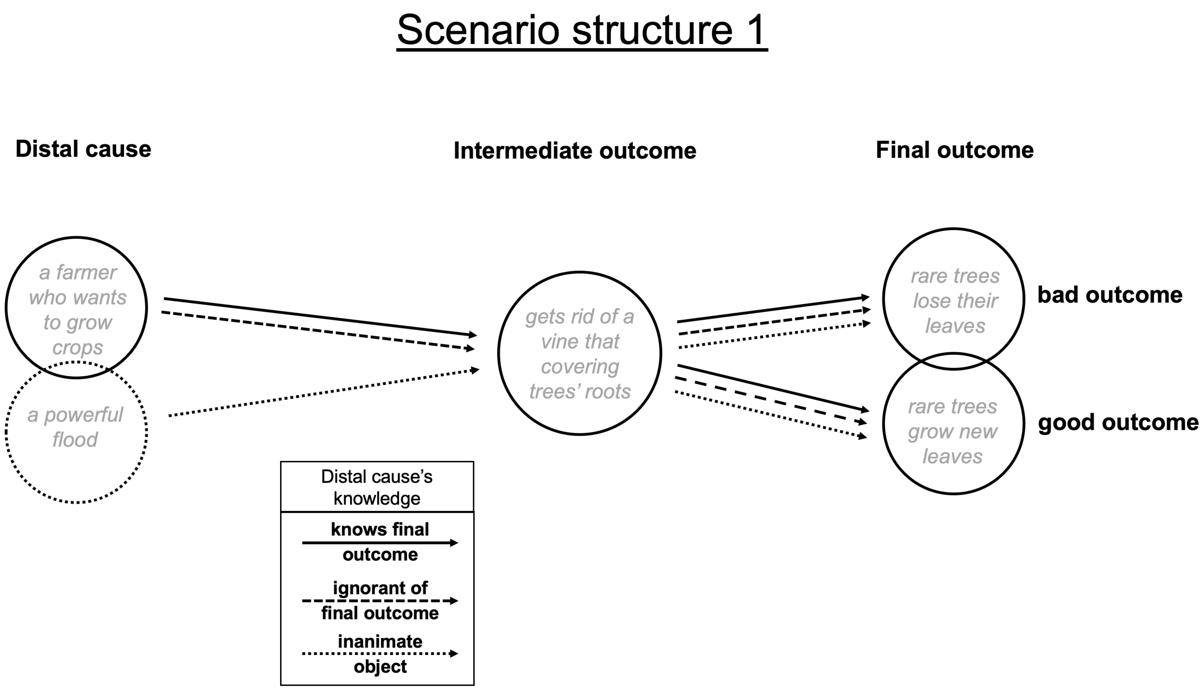


Figure 1: Study design for Experiment 1.

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. 92 participants (M_{age}=36.73, SD_{age}=12.21; 47 females) finished the whole study.

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 the most relevant counterfactual (36%), 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 (54%), 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 (48 %) than for good outcomes (56%), z = 4.74, $p < .001$. More importantly, we again observed a significant *Agent* × *Object* 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 (67%), than for good outcomes (41%), $z = 7.06$, $p < .001$. 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 (39%) or good outcomes (34%), z = 2.09, p = 0.292. That remains true for non-agentic objects in bad outcomes (52%) 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 = 0.66, $p < 0.001$) (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 -0.11, (95% CI = [-0.18, -0.04], p = 0.002). Controlling for the main effect of agent and knowledge, the proportion mediated is 0.307 (95% CI = [0.14, 0.54], p = 0.002).

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

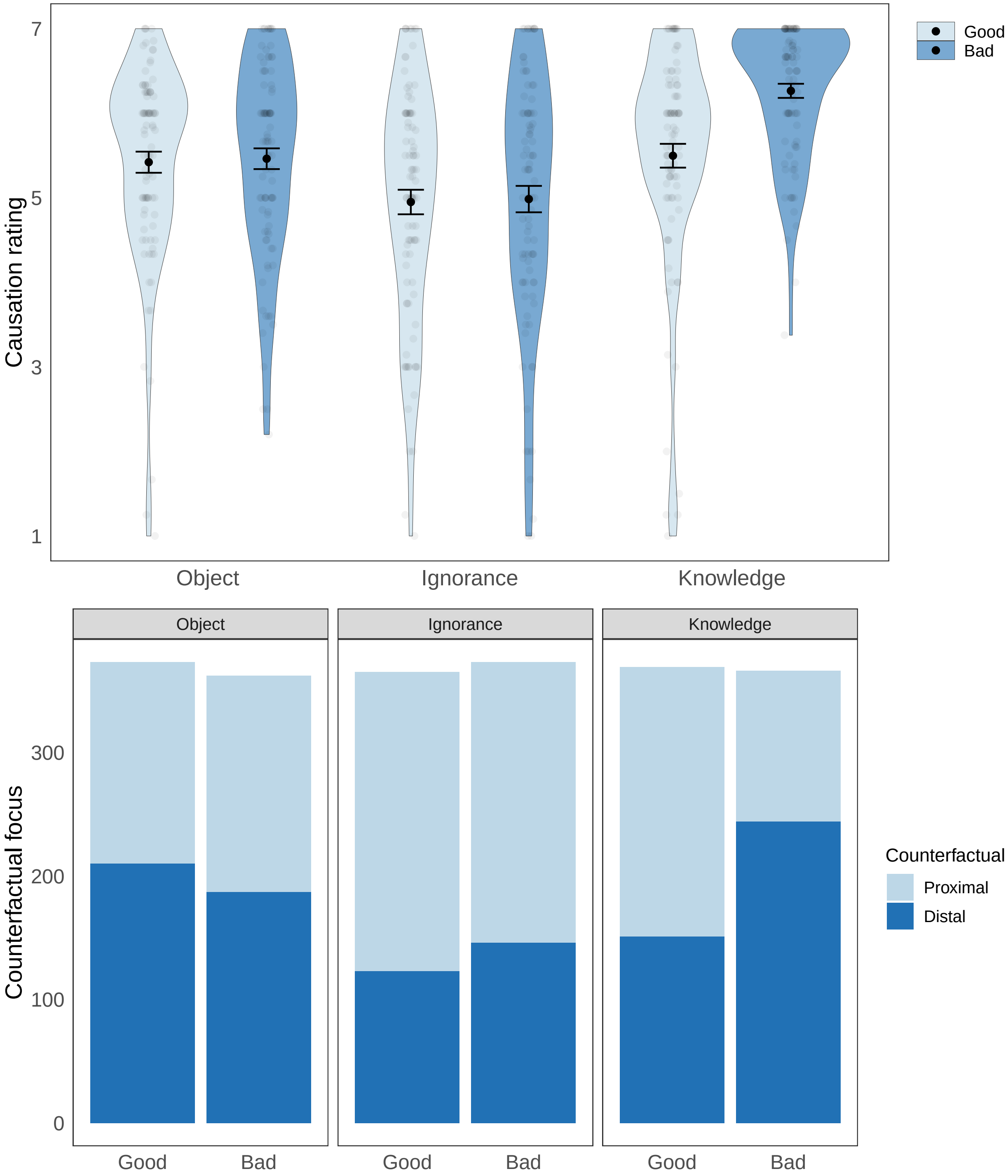


Figure 2: Participants' causal and counterfactual judgments as a function of both the kind of agent who initiated the causal chain and the valence of the outcome that eventuated. Error bars indicate +/- 1 *SEM*.

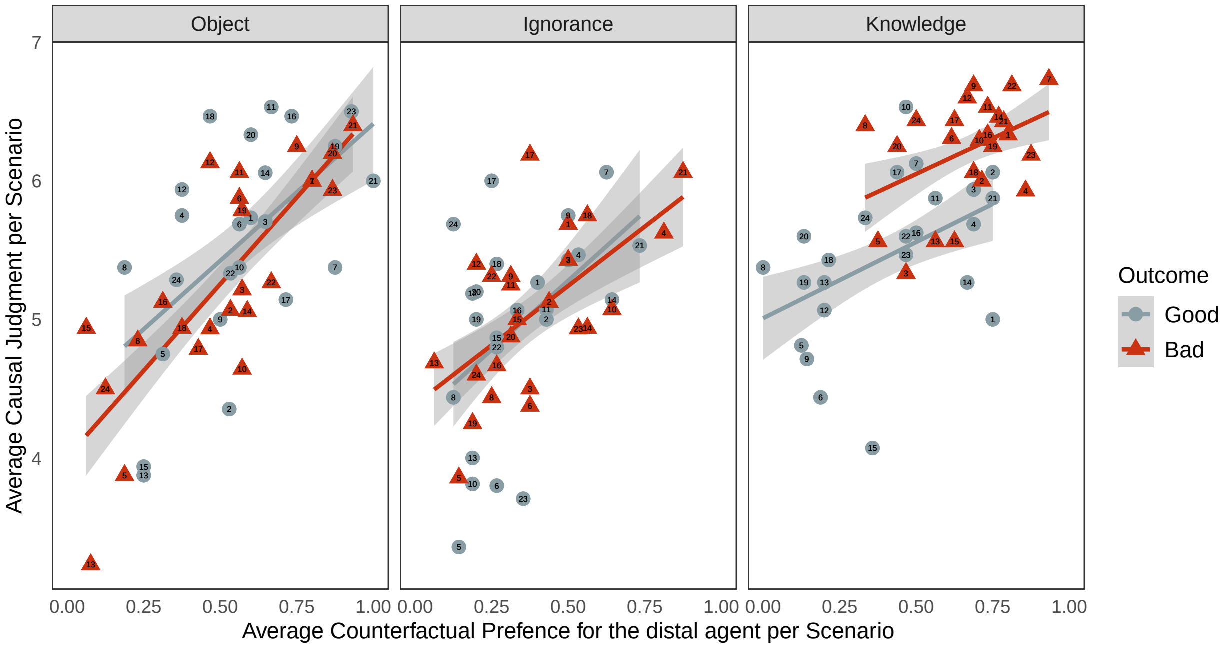


Figure 3: Depiction of the relationship between participants' causal and counterfactual judgments for each of the scenarios. Shape depicts the valence of the outcome; color indicates the kind of agent involved in the distal event; and the number indicates which of the 24 vignettes the judgments were 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. A similar pattern was also observed for inanimate objects which initiated the causal chain.

The interaction effect observed 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 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, or for inanimate objects. More generally, we also found that participants' causal and counterfactual judgments were tightly correlated in each of these conditions.

This experiment helps to establish the effect of norm violations in causal chains across many scenarios by manipulating what information the agent had access to. Interestingly, inanimate objects were judged in a somewhat similar way as ignorant human agents, suggesting that knowledge is indeed a factor in both causal judgment and counterfactual cognition. However, the experiment suffers from a confound that persists throughout the empirical work on this topic: the morality of the agent's action is always confounded with the goodness or badness of the outcome (see Hitchcock & Knobe (2009) and Kominsky et al. (2015) for two exceptions). In the next series of studies, we systematically and independently vary both the valence of the outcome and the morality of the agents' actions and ask how they contribute to participants' causal judgments independently.

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

In Experiment series 2, we innovate in two areas: first, by studying both types of prescriptive norms, moral norm and rational norm; second, by separating primary and secondary outcomes. The motivation is to expand the scope of causation attribution so that it broadens the scope beyond fixation on moral violations and addresses the challenge raised by researchers who argue that causal selection is driven disproportionately by the immorality of actions. Additionally, we aim to draw a distinction between (prescriptive) norm violations and valence of outcomes, concepts that have often been conflated in existing studies, by holding fixed the secondary outcome as negative and always having the agent be ignorant of the fact that this outcome would occur when acting.

In order to do so and to address the limitations of Experiment 1, we teased apart the prescriptive normality of the agent's action and the valence of outcome by having the action eventuating in two separate outcomes: one results in harm to the agent her-/himself or others, the other results in an objectively good or bad outcome as Experiment 1. Embedded in this design is our attempt to test whether the pattern of causal judgment remains the same between scenarios where a rational norm was violated and those where a moral norm was violated. As aforementioned, we set out with the goal to study both types of prescriptive norms: rational ane moral. More specifically, rational norm violation occurs when an agent performs an action that is not in her/his benefit while a moral norm violation occurs when an agent performs an action that brings harm to someone else.

The main causal chain in Experiment series 2 remains the same as the one in Experiment 1: we used scenarios where an agent (i.e., the distal cause) performed an action that became the proximal cause for a secondary outcome. On top of this, agents in Experiment series 2 would at the same time, either knowingly or ignorantly, bring about harm to others or themselves. The distal cause thus could be a knowing agent or an ignorant agent for this detrimental primary outcome. We then asked for participants' causal judgments on the *secondary outcome*. For Experiment 2a particularly, we kept the valence of the secondary outcome to always be negative. The overall study design described here is summarized in *Fig.* 4.

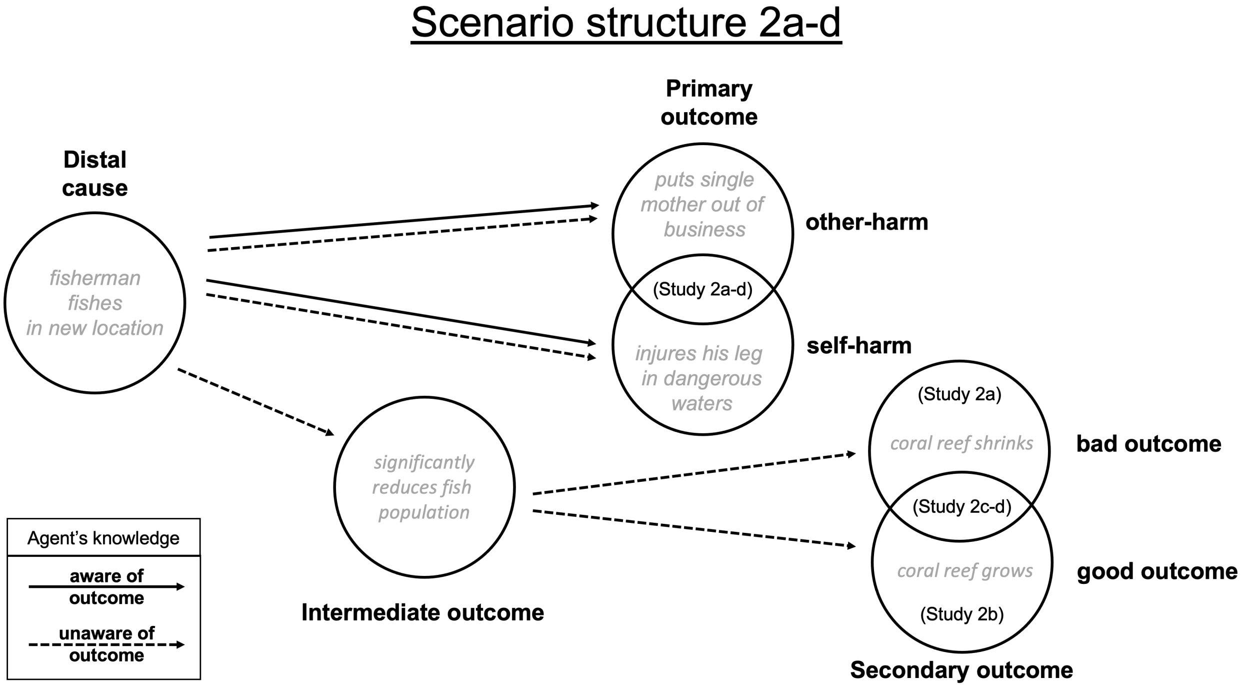


Figure 4: Study design for Experiment series 2

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) × 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 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 between harm to self ($M = 4.49$, 95% CI = [4.07, 4.91]) and harm to others ($M = 4.95$, 95% CI = [4.52, 5.38]), $t(15) = 0.12$, $p = 0.907$. 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 two significant insights that have not been widely documented in the literature: first, agents who knowingly preformed actions that were detrimental to selves or others were judged as more causal of the subsequent negative outcome, even when the immorality or irrationality of the action was completely independent of the eventual secondary outcome, while causal judgments about agents who were ignorant that their action did any harm saw no such effect; second, for causation, what matters is knowledge about norm violation in primary outcomes, not knowledge about valence of secondary outcomes.

That moral norm violation does not render a different effect from rational norm violation also implies that the underlying mechanism for causal selection may be the same for different types of prescriptive norms. Next, we went on to test if the findings apply in cases where the secondary outcomes are good in valence.

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

Some previous work has found an effect of moral norm violations even in good outcomes, while others argue that this effect is attenuated or even nonexistent (e.g., Alicke, Rose & Bloom, 2011). To investigate this exact question – whether the effect we see in Experiment 2a manifests in positive outcomes as well, such that knowledgeable agents are ascribed more causation than ignorant ones – we designed Experiment 2b which differs from Experiment 2a in only one way: the secondary outcome resulted from the intermediate event was always positive in this study. Again, we asked for participants' causal judgments on the *secondary outcome*.

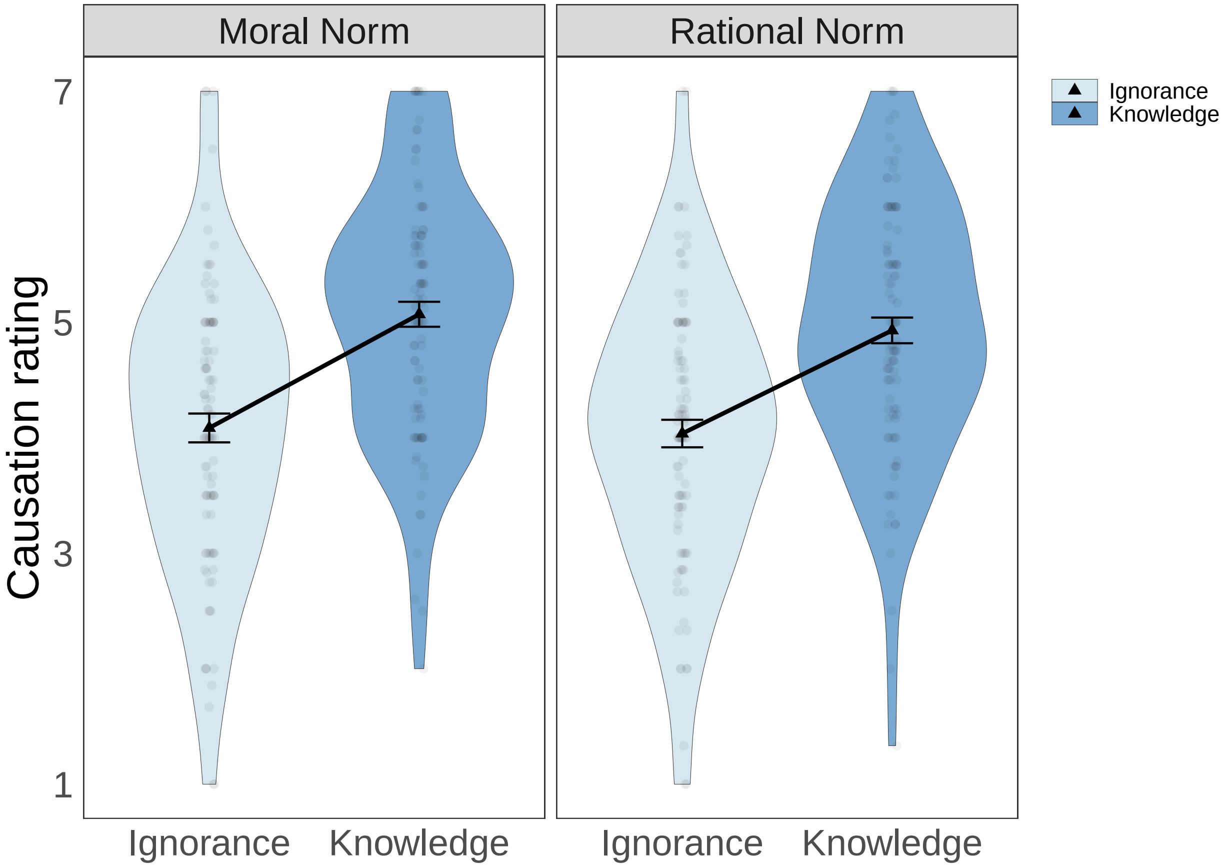


Figure 5: Participants' causal judgments as a function of both the knowledge of the agent who initiated the causal chain and the kind of norm that governed the agent's action, for both bad outcomes. Error bars indicate +/- 1 *SEM*.

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* × *Norm* 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.

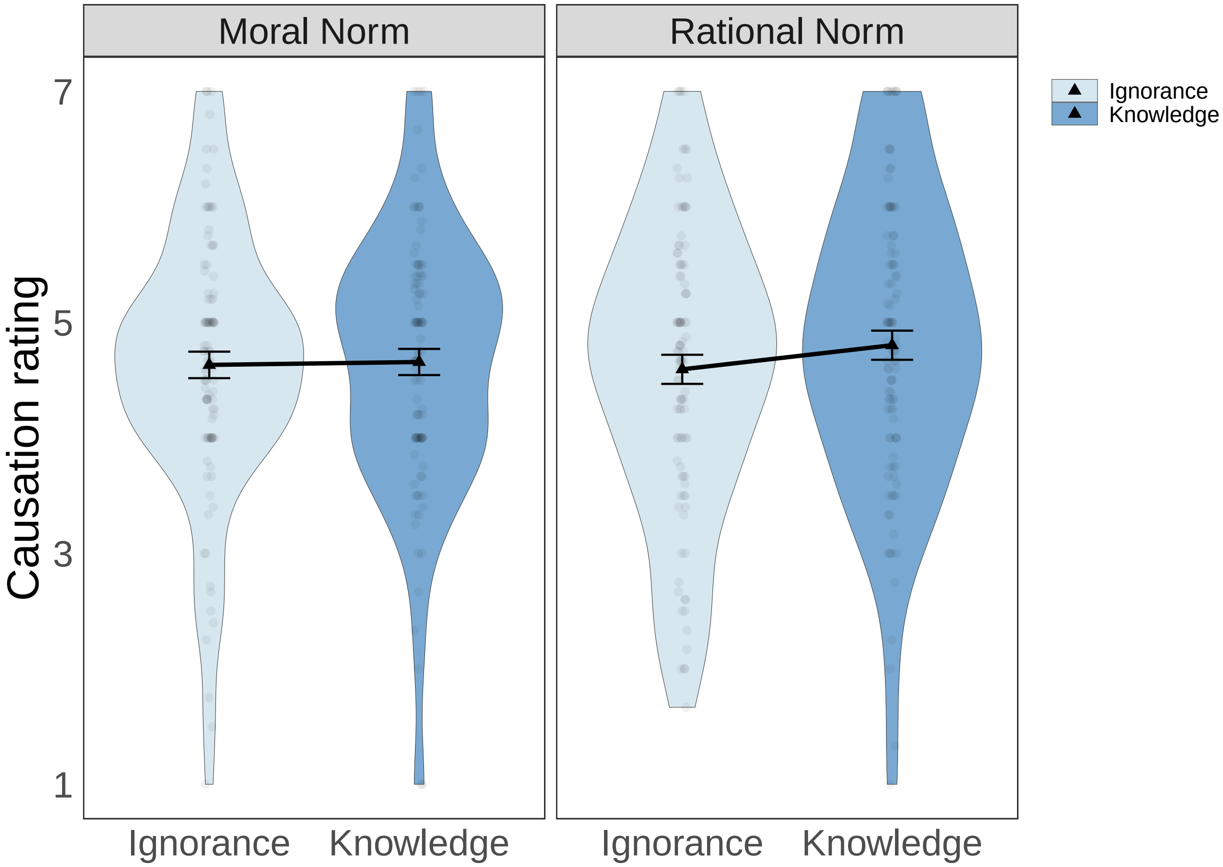


Figure 6: Participants' causal judgments as a function of both the knowledge of the agent who initiated the causal chain and the kind of norm that governed the agent's action, for good outcomes. Error bars indicate +/- 1 *SEM.*

Discussion

Compared to when the outcome was negative, in cases where secondary outcomes were positive, causal judgment was not affected by the type of prescriptive norm violated or whether the causing agent is informed about the consequence. Taken together, Experiment 2a and 2b suggest that agents' knowledge factors into causal attribution only when the secondary outcome is negative. In the upcoming study, we will be exploring this interaction by directly varying outcome valence.

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. Now, we would like to replicate these findings in a within-subjects design as we systematically manipulated the outcome valence. Based on prior results, we expected to see in varied outcomes an interaction effect between knowledge status and secondary outcome valence. In the current Experiment 2c, study design differs from that of 2a and 2b in one way only: the outcome may be positive or negative (see *Fig.* 4).

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 separate 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) × 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 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) = -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, 5.2]) and those who did so obliviously (M = 4.72, 95% CI = [4.42, 5.01]), t(15) = -2.73, p = 0.066 (*Fig.* 7). There was no significant *Knowledge* × *Norm* × *Outcome* interaction effect, $\chi^2(1)$ = 0.24, p = 0.625, *Norm* × *Outcome* interaction effect, $\chi^2(1)$ = 3.17, p = 0.075, or *Norm* × *Knowledge* interaction effect, $\chi^2(1)$ = 0, p = 0.963.

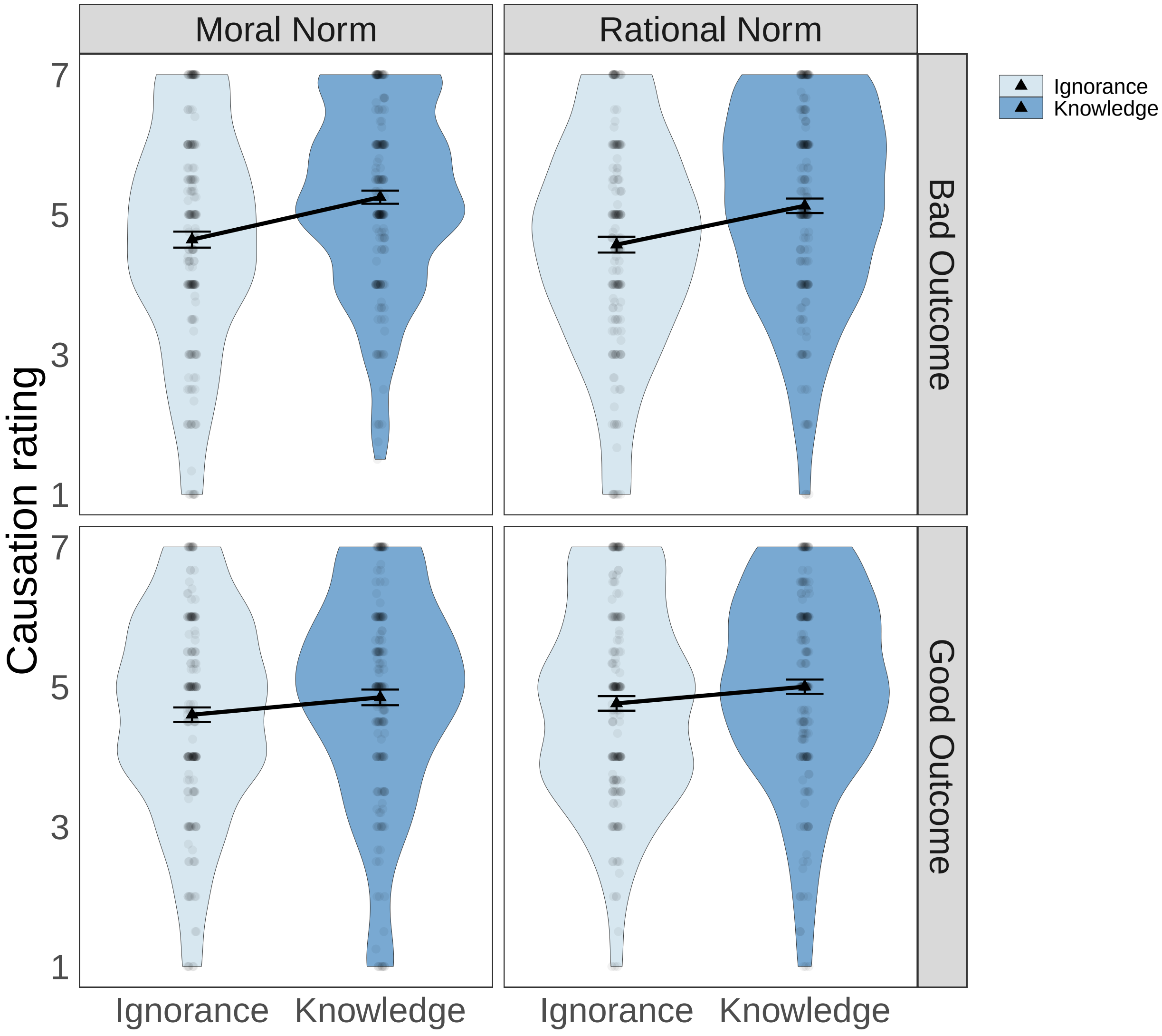


Figure 7: Participants' causal judgments as a function of both the knowledge of the agent who initiated the causal chain and the kind of norm that governed the agent's action, for both bad outcomes (left) and good outcomes (right). Error bars indicate +/- 1 *SEM*.

Discussion

Besides replicating the knowledge effect, the most substantial finding of Experiment 2c is the interaction between knowledge and outcome, such that more causation was attributed to knowledgeable agents than ignorant ones only when the secondary outcome was negative rather than positive. This interaction occurs in both moral norm violations (i.e., harming others) and rational norm violations (i.e., harming selves), suggesting that the it is unlikely to be explained by polysemy or the motivation to judge someone harshly because she/he did something bad. Put differently, if participants were confused by the multiple meanings embedded in the word "cause" or were inclined to have a lower opinion about someone who harmed others, we would see a differentiation in causal judgment between moral norm violation and rational norm violation. Therefore, we can say that causal selection works in a similar manner across different types of prescriptive norms.

However, if not explained by confusion or moral aversion, what is the underlying mechanism for this interaction effect? One previously undiscussed feature of outcome valence is that it is often tied with normality: good outcomes are perceived as more normal than bad outcomes (Bear & Knobe, 2017). In studies conducted so far, positive outcomes have been generally more descriptively likely than negative outcomes. We need to further dissect the phenomenon by disentangling descriptive normality and prescriptive normality, as we will do shortly (in Experiment series 3).

Experiment 2d: Combined analyses and new ratings on morality and rationality

checking variance by asking for morality and rationality ratings counterfactual: to see what's driving this effect

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) × 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.

Procedure

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’).

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

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 asked whether participants’ judgments of the morality and rationality of the agents’ actions tracked our manipulations as intended. Then we investigated the counterfactual responses in a similar way.

Morality Question

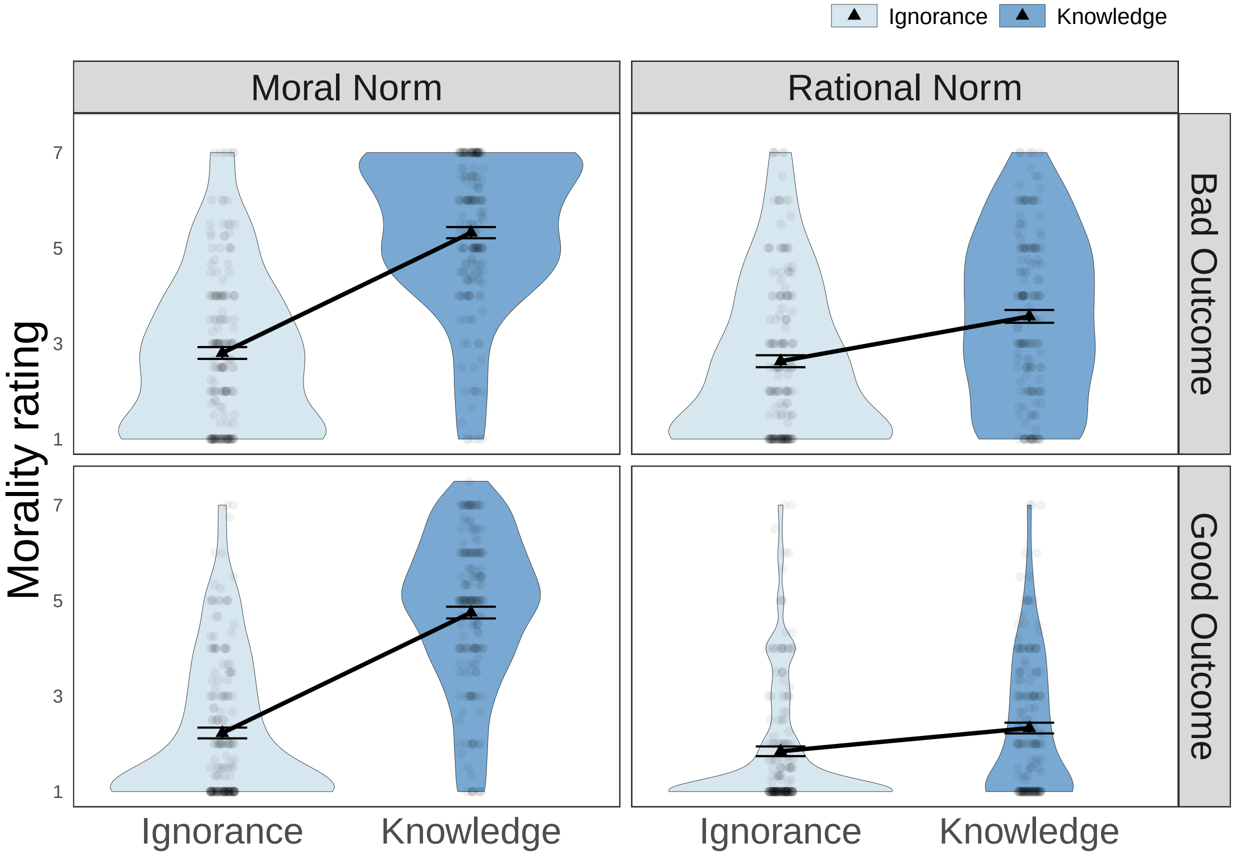
We started with participants’ moral judgments and found main effects across *Knowledge*, *Harm Type* and *Outcome* respectively. First, the main effect of the agent’s *Knowledge* about the consequence of their action, $\chi^2(1) = 0$, p <.001, indicates that events knowledgeable agents were overall seen 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]).

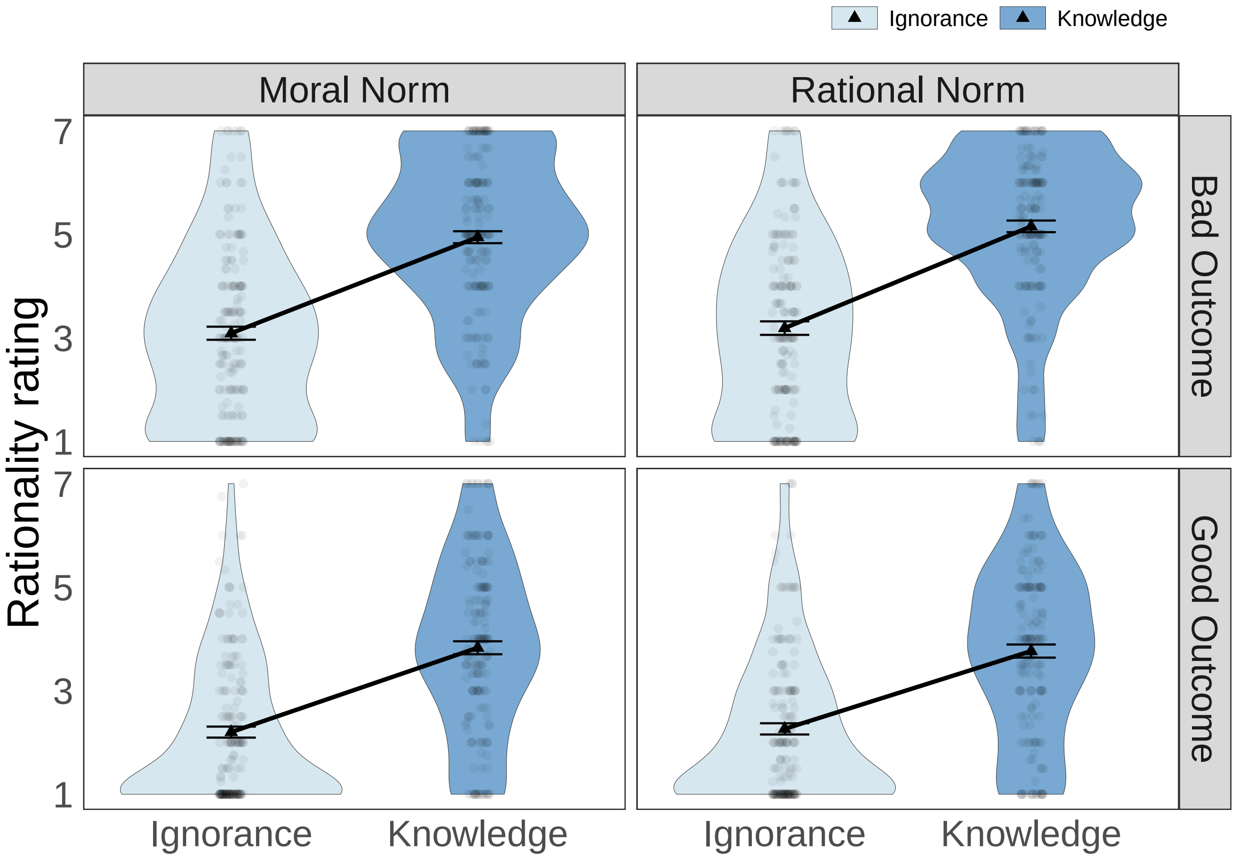
We then looked into interaction effects for moral ratings. A significant *Norm* × *Knowledge* interaction $\chi^2(1)$ = 260.1, p <.001, was observed. More specifically, when agents acted ignorantly, moral judgment on them harming selves (M = 5.75, 95% CI = [5.41, 6.08) and that on harming others (M = 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 *Norm* × *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* × *Norm* × *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).

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

Pooled causal judgments

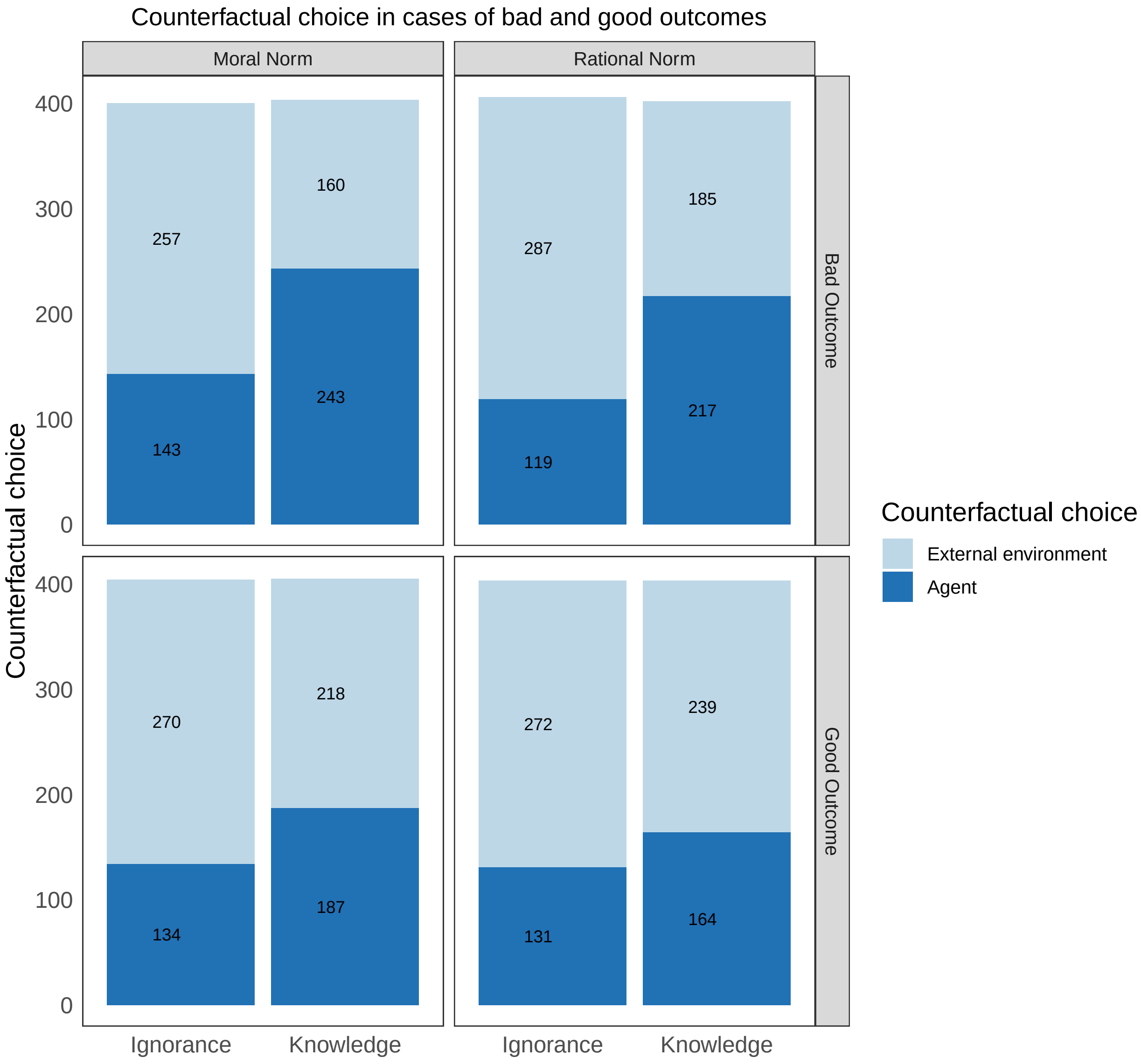


Figure 8: Counterfactual choices in good and bad outcomes.

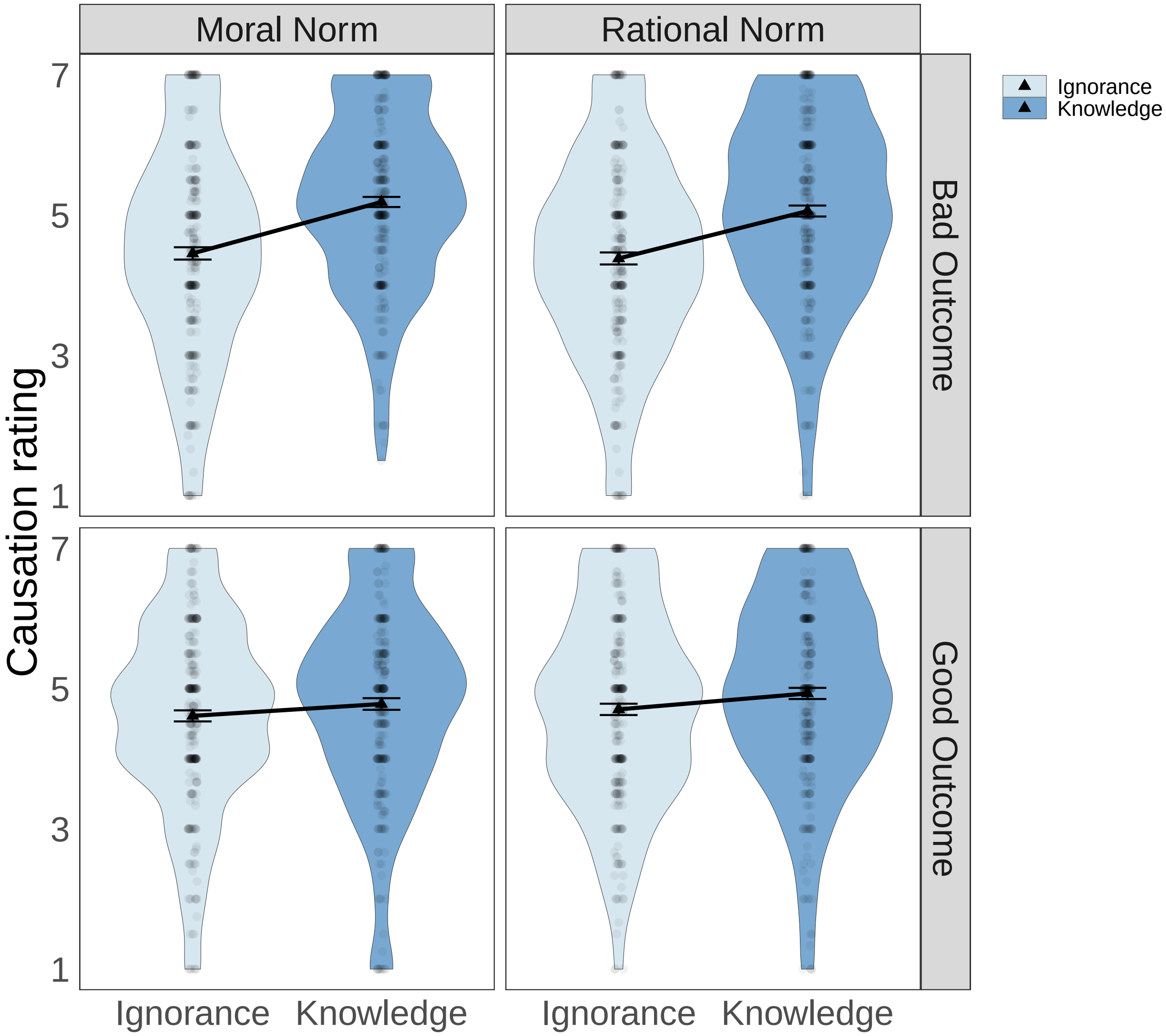


Figure 9: Participants' causal judgments as a function of knowledge status, outcome valence, and the kind of harm caused. Error bars indicate +/- 1 *SEM*.

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).

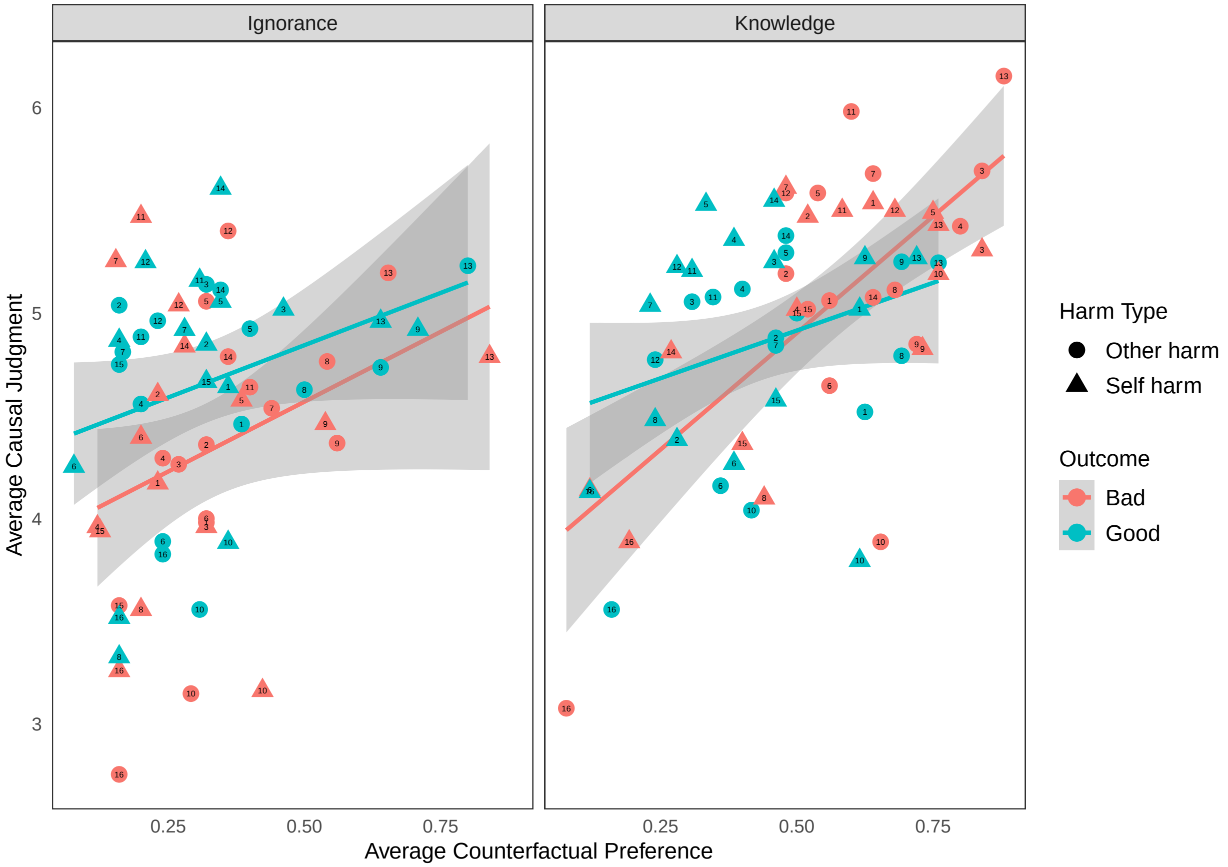


Figure 10: Depiction of the relationship between participants' causal and counterfactual judgments for each of the scenarios.

Appendix

Experiment 2d manipulation checks

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.