Immoral professors and malfunctioning tools:

Counterfactual relevance accounts explain the effect of norm violations on causal selection

Jonathan F. Kominsky*†

Rutgers University Newark, Department of Psychology

Jonathan Phillips*

Dartmouth College, Program in Cognitive Science,

Department of Psychological and Brain Sciences,

Department of Philosophy

*Authors contributed equally to this work

†Corresponding author

Address for reprints and correspondence	:	Jonathan F. Kominsky Psychology Department, Rutgers Newark 101 Warren St., Smith Hall Room 301 Newark, NJ 07102
Email	:	jonathan.kominsky@rutgers.edu
Phone/Fax	:	(617)-877-4412

Keywords: causation; norms; counterfactuals; morality; functional norms

Word count: 14944 (main text) + 485 (figure captions) + 1184 (tables and table captions)

Abstract

Causal judgments are widely known to be sensitive to violations of both prescriptive norms (e.g., immoral events) and statistical norms (e.g., improbable events). There is ongoing discussion as to whether both effects are best explained in a unified way through changes in the relevance of counterfactual possibilities, or whether these two effects arise from unrelated cognitive mechanisms. Recent work has shown that moral norm violations affect causal judgments of agents, but not inanimate artifacts used by those agents. These results have been interpreted as showing that prescriptive norm violations only affect causal reasoning about intentional agents, but not the use of inanimate artifacts, thereby providing evidence that the effect of prescriptive norm violations arises from mechanisms specific to reasoning about intentional agents, thereby casting doubt on a unified counterfactual analysis of causal reasoning. Four experiments explore this recent finding and provide clear support for a unified counterfactual analysis. Experiment 1 demonstrates that these newly observed patterns in causal judgments are closely mirrored by judgments of counterfactual relevance. Experiment 2 shows that the relationship between causal and counterfactual judgments is moderated by causal structure, as uniquely predicted by counterfactual accounts. Experiment 3 directly manipulated the relevance of counterfactual alternatives and finds that causal judgments of intentional agents and inanimate artifacts are similarly affected. Finally, Experiment 4 shows that prescriptive norm violations (in which artifacts *malfunction*) affect causal judgments of *inanimate artifacts* in much the same way that prescriptive norm violations (in which agents *act immorally*) affect causal judgments of *intentional agents*.

Keywords: causation; norms; counterfactuals; morality; functional norms

1. Introduction

A central question in research on causal cognition concerns the role of norms. It is well-known that both statistical and moral norms influence judgments of actual causation (i.e., a judgment that some particular event, *e*, was the cause of some particular outcome, *o*) (Alicke, 2000; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2015; Hitchcock & Knobe, 2009; Kominsky, Phillips, Gerstenberg, Lagnado & Knobe, 2015; Morris, Phillips, Gerstenberg, & Cushman, 2019; Morris, Phillips, Gerstenberg, Icard, Knobe, & Cushman, 2019). Specifically, when some outcome *o* depends on the occurrence of a set of antecedent events, e_1-e_n, people are more inclined to select a given antecedent event e_i as *the* cause of *o* if e_i was either very unlikely to happen or morally prohibited. Despite the widespread agreement on the existence of the phenomenon, there has been little corresponding agreement about how these effects should be explained.

Most researchers take the impact of *statistical* norms on causal judgments to reveal part of the basic underlying processes that support causal reasoning (e.g., Alicke, Rose, & Bloom 2011; Gerstenberg et al., 2015; Icard, Kominsky, & Knobe, 2017; Morris et al., 2019; Morris, et al., 2019; Samland & Waldmann, 2016). They differ, however, in whether they treat the impact of moral norms on causal judgments as arising from the same underlying processes or argue that it arises from a fundamentally different source.

On one side, a number of researchers have argued that the impact of both statistical and moral norms is best explained by changes in the *relevance* of counterfactual possibilities.[1] These counterfactual relevance accounts (CFR accounts hereafter) propose that when a norm violation

[1] These theories build on the general connection between causal and counterfactual reasoning (Hume, 1748; Lewis, 1973; Pearl, 2000) but additionally propose further ways of understanding what determines whether or not a given counterfactual will be relevant.

occurs, it increases the relevance of counterfactual alternatives wherein the norm violations are replaced by norm-conforming events (e.g., Halpern & Hitchcock, 2014; Icard et al., 2017). Causal judgments are then influenced by the necessity and/or sufficiency of each candidate cause in these cases. For example, if the outcome of the event does not occur in the counterfactual alternative in which the norm-violating action is transformed into a norm-conforming action, then the norm violation is typically judged to be more of a cause of that outcome, since the relevant counterfactuals highlight the fact that the norm violation is necessary for the outcome (e.g., Hitchcock & Knobe, 2009). In support of this account, recent work demonstrated that norm violations affect explicit assessments of counterfactual relevance in the same way that they affect causal judgments (Phillips, Luguri, & Knobe, 2015).

On the other side, other researchers have proposed that the impact of moral norms violations is not the same as the impact of other kinds of norm violations (e.g., Alicke et al., 2011; Sytsma, Livengood, & Rose, 2012). One recent account has suggested that the term "cause" is polysemous: It can be used to talk about whether some event causally contributed to an outcome, or it can be used to talk about whether an agent is morally responsible for an outcome (Samland & Waldmann, 2016). On this view, the effect of statistical norm violations is taken to arise from the ordinary processes involved in counterfactual cognition, while the impact of violations of moral norms is instead explained as part of *moral*, not causal, cognition: Participants are more likely to interpret the word "cause" as meaning "morally responsible" in cases where moral norms have been violated.

In support of this way of accounting for the impact of moral norms, Samland and Waldmann (2016) reported two important new observations: First, the violation of moral norms selectively influences causal judgments about whether *agents* caused an outcome, but not causal

judgments of whether the agents' use of *inanimate artifacts* caused that outcome. Second, factors that affect the moral responsibility of the norm violating agent (such as their knowledge states) also affect causal judgments (see also Samland, Josephs, Waldmann, & Rakoczy, 2016). These findings were taken to show that the changes in 'causal' judgments that tracked agents' moral responsibility are not genuinely reflecting intuitions about 'actual causation' (e.g., Danks, Rose, & Machery, 2014).

In this paper, we demonstrate that counterfactual relevance accounts have little trouble accounting for these new observations and further predict related phenomena that morality-specific accounts cannot explain. Thus, even setting considerations of parsimony aside, there is no reason to abandon a unified counterfactual account.

Our argument rests on four new kinds of evidence:

(1) We demonstrate that the pattern of causal judgments that has been interpreted as evidence against CFR accounts is in fact compatible with them: Participants' judgments of counterfactual relevance nearly perfectly predict their causal judgments in the cases that were assumed to be problematic.

(2) We show that, in cases where the outcome does *not* counterfactually depend on the norm violating event (cases of overdetermination), judgments of counterfactual relevance dissociate from causal judgments but continue to track moral norm violations, a prediction that is unique to CFR accounts; the observed pattern cannot be explained by morality-specific accounts.

(3) We *directly* manipulate the availability of relevant counterfactual alternatives in cases that do not involve norm violations and find that causal judgments of both agents and inanimate artifacts are affected by such experimental manipulations, demonstrating a causal relationship between counterfactual relevance and causal judgments for both agents and artifacts.

(4) Finally, we establish that prescriptive norm violations *do* affect causal judgments of inanimate artifacts when the norm being violated actually concerns the artifacts themselves (i.e., a norm of proper functioning).

Before reporting these four new pieces of evidence, we elaborate on two important aspects of CFR accounts that have been underspecified in previous work, and then explain how each of our four experiments help answer a key questions at the center of the debate over the impact of norms on causal judgments.

1.1. What counts as a norm violation?

The key idea of CFR accounts is that norm violations lead people to consider counterfactual alternatives that are more "normal" (Kahneman & Miller, 1986; Phillips & Knobe, 2018), and that people evaluate whether the outcome would still obtain in these counterfactual possibilities (Hitchcock & Knobe, 2009; Icard et al., 2017). Importantly, the content of the relevant counterfactual alternatives reflects the nature of the norm being violated. If an agent commits a moral norm violation, then the relevant counterfactual alternatives are those in which the agent conforms to the moral norm rather than violating it. Then, the question is simply whether, in these counterfactual possibilities, the change in the agent's actions results in a corresponding change in the occurrence of the outcome; if so, then the agent should tend to be seen as more of a cause of the outcome.

Previous work on CFR accounts has not tried to offer a clear definition of what constitutes a prescriptive or moral norm violation. Instead, this work has largely proceeded by taking intuitively clear examples of prescriptive moral norm violations and then characterizing the effect that these kinds of norm violations have on causal judgments (e.g., Hitchcock &

Knobe, 2009; Kominsky et al., 2015). An unfortunate consequence of this approach has been that there are many cases where it remains unclear whether or not a norm has been violated, and thus also unclear whether CFR accounts should or should not predict that causal judgments will be affected.

To see this, consider two broad ways of understanding when an event will count as an instance of a moral norm violation. One possible interpretation is that the relevant notion of 'norm' can simply be reduced to violations of a set of known prescriptive rules that govern a given situation (e.g., 'thou shalt not kill' or 'do not eat the cookies'). On this interpretation, a norm will be violated in any event in which someone is killed, or a cookie is eaten. This interpretation of norm violations broadly aligns with the one assumed by Samland and Waldmann (2016).

An important alternative way of understanding norm violations holds that moral norm violations cannot be so easily reduced to instances where a behavior violates a prescriptive rule. This alternative understanding is typical in work on norms in philosophy and is nicely captured by the *Stanford Encyclopedia of Philosophy* article on social norms:

> In a nutshell, norms refer to actions over which people have control, and are supported by shared expectations about what should or should not be done in different types of social situations. However, norms cannot be identified just with observable behavior, nor can they merely be equated with normative beliefs. (Bicchieri, Muldoon, & Sontuoso, 2018, 8)

Instead of focusing on when a behavior could be classified as a violation of a given prescriptive rule, this approach instead focuses on normative expectations of the agent, that is, what an agent

should or should not do given their beliefs (see Prentice, 2007, for a similar approach to understanding prescriptive norms in social psychology).

A critical example where these two interpretations make different predictions are cases in which an agent breaks an established rule but lacked some relevant piece of knowledge when acting. These are the kinds of cases tested in Samland & Waldmann's Experiment 4, in which they manipulated whether an agent violated a norm deliberately, unintentionally, ignorantly, or because they were deceived. On the first, purely behavior-based way of interpreting moral norm violations, this event clearly counts as a norm violation in all four cases because there is a prescriptive rule prohibiting that action. On the second, expectation-based interpretation, however, the event should not count as an instance of a moral violation except when it was done deliberately, since there is no general expectation that agents should abide by rules even when they lack critical knowledge of them.

While previous work on counterfactual relevance accounts have primarily sought to characterize the effect of prescriptive norm violations on causal judgments, an important further step is to better circumscribe what does and does not count as a norm violation. Following the traditional approach in philosophy and psychology, we propose that prescriptive norm violations ought to be understood as instances in which there is a violation of general expectations about what should have occurred (e.g., what that agent should have done). Accordingly, we predict that participants judgments of counterfactual relevance will track this understanding of norm violations rather than simply instances in which a prescriptive rule is violated. We test this in Experiment 1 using the mental state manipulations created by Samland and Waldmann (2016).

1.2. How are counterfactual events generated?

Samland and Waldmann (2016) found that moral violations affected causal judgments of the norm-violating agent but not causal judgments of their use of an inanimate tool, and argued that CFR accounts are committed to predicting a different pattern. In particular, they suggested that the moral violation should lead people to consider counterfactual alternatives involving differences in the agents' actions, and so questions referring to the agent, the action, or their use of an artifact should all be equally affected by the moral norm violation (p. 171). Based on their finding that causal judgments of the *agent* are primarily affected, Samland and Waldmann conclude that these cases represent an instance in which causal judgments clearly dissociate from the relevant counterfactuals. However, they do not empirically examine whether participants found it relevant to consider counterfactual alternatives involving the agent, the agent's action, or their use of the artifact. Accordingly, the putative dissociation hinges on the predictions they attribute to the specific CFR account they are responding to (that initially offered by Hitchcock & Knobe, 2009).

The way in which counterfactual possibilities are generated has been left to intuition in existing work, and so understanding whether the patterns of responses identified by Samland and Waldmann (2016) can be explained by CFR accounts requires further elaboration of existing theories. Consider two possible views:

One way that counterfactuals could be generated is that people consider counterfactuals in which most aspects of the norm-violating event are represented and mutated. If right, then when a norm violating event occurs involving an agent using some artifact, it should become more relevant to consider how the agent herself could have been different, but also ways the agent's action could have been altered, and ways the functioning of the artifacts they used could

have been changed. After all, the relevant counterfactuals may involve changes to all of these aspects of the norm-violating event, since replacing the agent's action will necessarily change the functioning of the artifacts they used.

An alternative possibility, which we think is more realistic, is that when a norm violating event occurs, people do not generate counterfactuals that involve representing or mutating all of these aspects of the event. Instead, they represent the event more granularly, focusing specifically on aspects of the event directly involved in the norm violation. Thus, for example, when a driver violates a norm by not obeying the signs on the road, people might only consider ways that the driver could have intentionally acted differently or made a different decision, e.g., she could have decided *not* to disobey the 'no left turn' sign, but not ways in which that action could have *been performed* or *gone* differently (e.g., her hands could have *not* moved counter-clockwise), or ways in which the tool used could have *functioned differently* (e.g., the front wheels of the car could have *not* reoriented toward the left when the driver turned the wheel). If this alternative is right, then when participants are asked a causal question about an agent, people's judgments are likely to be based primarily on whether it is relevant to consider counterfactual possibilities in which the *agent* made a different decision and performed a different action, and not on counterfactuals involving changes to other aspects of the event (e.g., the way the artifact functioned during use). Our studies provide an empirical test of this hypothesis.

1.3. Four unanswered questions addressed by the present experiments.

With these theoretical clarifications in mind, we can now pose four questions that our experiments set out to answer.

1.3.1. Which counterfactuals are relevant? Our proposed elaboration of existing CFR accounts breaks this broader question into two related, and testable, issues:

(1) Do expectation-based or behavior-based accounts of prescriptive norms better account for participants' judgments of counterfactual relevance?

(2) Do norm violations lead people to regard as relevant counterfactual alternatives that encompass many aspects of the norm-violating event, or do they only lead people to regard as relevant counterfactual alternatives that focus more granularly on the norm violating event (e.g., the agent's decision in a moral norm violation)?

We answer both of these questions using the scenario in Experiment 4 in Samland and Waldmann (2016), by adding matched judgments of counterfactual relevance: "how relevant is it to consider how the [agent/the agent's action/the artifact used] could have been different?"

Considering the first question, a purely behavior-based theory of moral norm violations holds that the norm violation should be tied to the action, regardless of the mental states of the actor. This account would therefore seem to predict that there should be no effect of mental state on counterfactual relevance judgments. In contrast, the expectation-based theories of norm-violations predict that changes in an agent's mental states will directly affect whether or not the agent is perceived as having violated a norm, and consequently whether or not it is relevant to consider counterfactual alternatives focused on the agent. This second account, which we favor, therefore predicts that counterfactual relevance judgments should be affected by mental state manipulations, such that counterfactuals for an agent's unintentional, ignorant, or deceived norm violations should be judged less relevant than counterfactuals for intentional violations.

Second, if people generate counterfactuals in a way that richly represents and changes many aspects of the relevant event, then judgments of counterfactual relevance of the agent's

action and the artifact used by the agent should track each other (either they will both be relevant in the case of a norm violation, or neither will if no norm is violated). This pattern, if observed, would result in a dissociation between counterfactual relevance and causal judgments, since only causal judgments of the agent but not the artifact were affected by moral norm violations. In contrast, the account we favor on which people more granularly generate counterfactual alternatives predicts an increase for the relevance of alternatives focusing on the norm-violating agent but no corresponding increase in the relevance of alternatives focusing on the artifact used. If correct, we would *not* see a dissociation between counterfactual relevance and causal judgments.

In short, both of the ways in which we have made CFR accounts more precise make the falsifiable prediction that counterfactual relevance judgments should closely correlate with causal judgments across the various scenarios used by Samland and Waldmann (2016). We test this prediction in Experiment 1.

1.3.2. When are counterfactuals relevant? As indicated above, CFR accounts have consistently proposed that causal judgments and counterfactual relevance judgments will often be closely related to one another. However, in cases of moral norm violations, a natural interpretation of this correlation is that it arises from some third variable that affects both causal and counterfactual judgments. For example, it may be that both causal and counterfactual questions are easily interpreted as being questions about moral responsibility. If this were the case, then it would not be surprising to observe that the two judgments are highly correlated, and more importantly, this correlation certainly would not be evidence that causal judgments depend on counterfactuals, as argued for by CFR accounts. Helpfully though, there are cases in which CFR accounts predict that the correlation between counterfactual relevance judgments and causal

judgments should break down. Specifically, Icard et al. (2017) argued that in *overdetermined* cases, when the norm violation is sufficient but not necessary to bring about the outcome, causal judgments of the norm violating event should *decrease* rather than increase. The basis of this prediction is that counterfactuals in which the norm violation does not occur are still *more* relevant, but in overdetermined cases, such counterfactual possibilities only serve to reinforce the fact that the outcome did not actually depend on the norm violating event occurring. In other words, when you consider a counterfactual possibility in which the norm violating event does not occur, you will be confronted with the fact that the outcome will still occur even if the norm violating event does not, since the outcome was overdetermined. Thus, increasing the consideration of these counterfactuals actually leads to reduced causal judgments.

If the pattern predicted by CFR accounts were observed, then it would provide clear evidence both that (i) causal questions are not simply being interpreted as questions about moral accountability, and (ii) judgments of counterfactual relevance only correlate closely with causal judgments when the outcome counterfactually depends on the norm violating event. We provide new data on this untested prediction in Experiment 2.

1.3.3. Do counterfactuals really cause changes in causal judgments? In Experiments 1 and 2, we manipulate the occurrence of norm violations and seek to demonstrate that the impact on causal and counterfactual judgments meets the predictions of CFR accounts. Accordingly, these experiments provide only correlational evidence. Yet, CFR accounts make the stronger causal claim that causal judgments depend on counterfactual judgments. To test this stronger hypothesis, we can temporarily put norm violations aside and seek to more directly manipulate the presence of relevant counterfactuals (Phillips et al., 2015). By doing so, we can test whether the specific content of the relevant counterfactual alternative causes the predicted change in

causal judgments. In Experiment 3a and 3b, we ask a group of participants to generate relevant alternatives to the way the artifact functioned, and then test whether this specifically affects their causal judgments of the artifact; at the same time, we ask another group of participants to instead generate relevant counterfactual alternatives to what the agent did and then test whether this primarily affects their causal judgments of the agent.

1.3.4. Are inanimate objects affected by prescriptive norm violations? By limiting our investigation to moral violations and direct manipulations of counterfactual relevance, we have not considered a key prediction of the elaborated CFR account: Causal judgments of artifacts should also be affected by prescriptive norm violations, but only when the prescriptive norm being violated actually governs the artifact. That is, when the normative expectations have to do with the *functioning* of the artifact (e.g., does it do what it was designed or expected to do?) rather than the intent of the person using it. Of course, artifacts cannot act immorally; as Samland and Waldmann (2016) rightly point out, only intentional agents can commit moral violations. Accordingly, our elaborated CFR account does not expect moral norm violations to affect the relevance of alternatives relating to the functioning of inanimate artifacts. However, there are other types of prescriptive norms which should affect the relevance of counterfactual alternatives relating to artifacts, namely, norms of *proper function*. Such norms have previously been found to affect causal judgments of artifacts (Hitchcock & Knobe, 2009; Livengood, Sytsma, & Rose, 2017). For example, Hitchcock and Knobe (2009) found that people judged that a red wire which touched a battery when it was not supposed to was more of a cause of a short-circuit than a black wire that was supposed to touch the battery. However, no studies (to our knowledge) have provided direct evidence that malfunctions change the relevance of counterfactual alternatives relating to artifacts. Furthermore, no previous work on malfunctions has looked at cases like the

ones used in Samland and Waldmann, in which different artifacts are used by different agents to produce some outcome.

Therefore, in Experiment 4, we introduced either a moral norm violation or a functional norm violation in a case where an agent uses an object and this action results in the occurrence of some outcome. We ask participants for both causal judgments and counterfactual relevance judgments of both the agents and the objects. This allows us to test the general prediction of CFR accounts: The moral violation should primarily affect counterfactual and therefore causal judgments of the agent and not the artifact used (as found in Experiment 1), while the functional norm violation should primarily affect counterfactual and therefore causal judgments of the artifact, but not the agent who used the artifact.

2. Experiment 1

In Experiment 1, we set out to replicate the pattern of causal judgments reported by Samland and Waldmann (2016) and compare that pattern to a measure of counterfactual relevance. Our stimuli were identical to Samland and Waldmann's stimuli in their Experiment 4. The only difference was that, prior to asking a causal judgment question, we asked "in thinking about how things could have happened differently, how relevant is it to consider the following", and asked participants to select either or both of the agents, to select either or both of the actions, or to select either or both of the artifacts (all of the stimuli and questions are available in the Supplemental Materials). Under Samland and Waldmann's polysemy account, there are no clear predictions for counterfactual relevance judgments. Under an elaborated CFR account, the counterfactual relevance of alternatives focusing on the agent should be affected by the moral violation, and should be closely correlated with causal judgments across conditions. In contrast,

the relevance of counterfactual alternatives focusing on the agent's behavior, or the use of the inanimate object should be less influenced by the moral violation, and thus continue to align with participants' causal judgments.

In addition, Samland and Waldmann found that causal judgments of the agents were strongly affected by the agents' mental states and take this as evidence against CFR accounts. According to their interpretation which draws on a behavior-based understanding of norms, the mental state manipulation should not affect counterfactual relevance judgments nor "true" causal judgments (and thus we should interpret the causal judgments they report as accountability judgments). Our elaborated CFR account presents a plausible alternative prediction: If a moral norm violation is instead understood as expectation-based, then the agent's mental state will obviously affect *whether the agent violated a norm at all*. This prediction is intuitive enough: If an agent knowingly decides to break a rule, then it seems clear enough that this agent did something wrong—something that violates people's general expectations of what the agent *should have done*. However, if another agent does the same thing but only does so because they were unfairly deceived about what the rule is, then it is no longer clear that the agent did something wrong; there is no general expectation that agents should not do actions that violate rules even when they are unfairly deceived into believing the rule did not exist (indeed in such cases the deception itself is the intentional moral violation). As such, our elaborated CFR account predicts that both counterfactual relevance and causal judgments should be affected in similar ways by the mental state manipulations in Samland and Waldmann (2016).

2.1. Methods.

2.1.1. Participants. 610 participants (M_{age} = 37.28, SD_{age} = 12.14; 338 female, 1 unreported) from Amazon Mechanical Turk participated for $0.25 in compensation. Participant recruitment was automated through TurkPrime (www.turkprime.com) to prevent repeat participation and limit recruitment to participants with a previously established high approval rate.

2.1.2. Stimuli and procedure. This experiment was an exact replication of Samland and Waldmann (2016)'s Experiment 4 except for the addition of a second DV. This experiment used a 4 (Norm violation: Standard, Unintended, Ignorant, Deceived) × 3 (Question: Agent, Action, Object) design, administered fully between-subjects. The study materials were presented in Qualtrics (Qualtrics, 2005).

Participants read one of four vignettes identical to those used in Samland and Waldmann (2016) (the full text can be found in the Supplementary Materials). In all conditions, a man named Tom owns a garden and has two gardeners, Alex and Benni, who each take care of 1/3 of the plants on their own and jointly tend to the remaining 1/3. Additionally, Alex and Benni always use two fertilizers "A-X200®" and "B-Y33®". Tom reads that fertilizers are good for plants, but using more than one kind of fertilizer could damage his plants by drying them out, so Tom decides he wants both gardeners to use only fertilizer A-X200. In all cases, Alex applies fertilizer A-X200, but Benni applies fertilizer B-Y33, and the plants cared for by both of them are damaged.

The four norm conditions varied the reason that Benni used B-Y33. In the *Standard norm-violation condition*, Benni simply decides to use B-Y33, intentionally defying Tom's instructions; in the *Unintended norm-violation condition*, Benni believed he was following Tom's instructions by applying A-X200, but accidentally applied B-Y33; in the *Ignorant norm-*

violation condition, Tom neglects to tell Benni to use only A-X200, and Benni happens to use B-Y33 instead; and in the *Deceived norm-violation condition*, Alex is supposed to convey Tom's instructions, but deliberately lies to Benni about which fertilizer he is supposed to use to get Benni in trouble.

Additionally, the questions that participants answered varied. As in Samland and Waldmann (2016), participants were either asked questions that focused on the two agents ("Alex" and "Benni"), the two actions ("the application of fertilizer by Alex" and "the application of fertilizer by Benni"), or the use of two chemicals ("the application of chemical A-X200" and "the application of chemical B-Y33").

After reading the vignette, participants were first asked whether it was *relevant* to consider counterfactual alternatives to some aspect of the event (following Phillips et al., 2015). For example, in the Agent condition, participants were asked whether it was relevant to consider what Alex(/Benni) could have done differently. Subsequently, as in Samland and Waldmann (2016), participants were asked to judge who or what caused the plants to dry up (appropriate to the Question condition) and were allowed to choose one or both of the two options.[2]

Following this question, participants received two check questions that tested their understanding of which chemicals were applied by which gardener and which chemicals Tom wanted each gardener to use. Following Samland and Waldmann (2016), they were also asked to estimate the proportion of the flowers that dried when (1) only fertilizer A-X200 was applied, (2) only fertilizer B-Y33 was applied, and (3) both were applied (see Supplementary Materials for replication). All data, stimuli, and analysis code are available at: https://osf.io/cp2d5.

[2] This fixed order of questions was used because Samland and Waldmann (2016) already established the pattern of causal judgments without first asking for judgments of counterfactual relevance, and thus any effect of first answering a question of counterfactual relevance could be detected by comparing causal judgments across the two studies.

2.2. Results.

As in Samland and Waldmann (2016), we excluded participants who did not answer both of the check questions correctly (171/610, or ~28%; for comparison, Samland and Waldmann excluded 285/869 or ~33%) and analyzed the remaining 439 participants' judgments.[3] Post-hoc power analyses showed that all statistical tests had power ≥ 99% to detect the observed effects.

In all analyses, we looked at the proportion of participants who had selected each option, i.e. the proportion of participants who selected Benni/Benni's action/Fertilizer B and the proportion who selected Alex/Alex's action/Fertilizer A were computed for each condition (the full pattern of responses can be found in Tables S1-2). Our primary analysis strategy was to look at two sets of correlations. First, we compared the causal judgments in our current experiment to those in Samland and Waldmann (2016) (who helpfully made their data publicly available), to ensure that our results replicated theirs. Second, within this experiment alone, we compared causal judgments and counterfactual relevance judgments at both the level of participant and the level of condition.

We both qualitatively and quantitatively replicated the pattern of causal judgments observed in Samland and Waldmann (2016). At the level of each condition, participants' average causal judgments in our experiment were highly correlated with those of Samland and Waldmann, $r = 0.950$ [95% CI 0.887, 0.979], $p < .001$. We graph this replication by comparing the average causal judgment for each of the two agents in each of the 12 conditions for both our data and Samland and Waldmann's data (Fig. 1, left panel, see also Table S3 for complete

[3] For each experiment we ran a chi-square test of exclusions across all between-subjects conditions to determine whether exclusion rate varied by condition. In all experiments this test showed no significant differences, *p*s > .1.

information on the replication of the key statistical tests reported in Samland and Waldmann, 2016).

Next, to examine the effects of the manipulations on both causation and counterfactual relevance judgments, we categorized participants' responses as assigning causal responsibility (or counterfactual relevance) to (1) only the norm-violating agent, (2) both agents, or (3) only the norm-conforming agent (these labels refer to Samland & Waldmann's behavior-based understanding of norms to make it easier to compare across studies). We then subjected both kinds of judgments to a proportional odds logistic regression using the probit function in the MASS package in R. For causal judgments, we observed an effect of the norm-condition (LRT = 20.49 [df = 3], $p < .001$), an effect of question (LRT = 44.53 [df = 2], $p < .001$), and critically, a Norm violation × Question interaction effect (LRT = 19.94 [df = 6], $p = .003$). For relevance judgments, we observed a highly similar pattern of results: an effect of Norm violation (LRT = 13.93 [df = 3], $p = .003$), an effect of Question (LRT = 73.34 [df = 2], $p < .001$), and a Norm violation × Question interaction effect (LRT = 14.15 [df = 6], $p = .028$). Further, at the level of each condition, participants' average causal judgments in our experiment were extremely similar to that of their judgments of counterfactual relevance, $r = 0.848$ [0.675, 0.932], $p < .001$. The similarity of the pattern of these two judgments across the various conditions can be seen in Fig. 1, right panel.

Moreover, at the level of each participant, their judgments of causal responsibility were highly correlated with their judgments of whether it was relevant to consider alternatives to the events. This was true both for judgments of the norm-violating agent/action/artifact, ($r = 0.553$ [.484, .615], $p < .001$) and for the norm-conforming agent/action/artifact ($r = 0.406$ [.325, .481], $p < .001$). Crucially, this relationship held whether participants were making judgments about

agents (r = 0.651 [.575, .715], p < .001), actions (r = 0.262 [.157, .362], p < .001), or the use of inanimate artifacts (r = 0.280 [.172, .382], p < .001).

----- Insert Figure 1 about here ----

2.3. Discussion.

We successfully replicated Samland and Waldmann (2016)'s results for causal judgments but additionally found the exact same pattern for counterfactual relevance judgments. First, counterfactual relevance judgments of agents were strongly affected by the agents' mental states, as predicted by our expectation-based account of what constitutes a "moral violation" (see §1.1 and §1.3.1). Critically, when the agent committed a moral violation, counterfactual relevance judgments for the agent increased, but counterfactual relevance judgments for their use of the artifact did not, demonstrating that CFR accounts are compatible with the difference in causal judgments for agents and the objects they use. Furthermore, we found a close correlation between causal and counterfactual relevance judgments both across individual judgments and averaged across conditions, even for non-agentic objects that obviously did not commit moral norm violations. Thus, this experiment shows that the results of Samland and Waldmann are compatible with both their polysemy account and an elaborated CFR account. In Experiment 2, we next consider two possible explanations for *why* we find such a robust relationship between judgments of counterfactual relevance and causal selection, and test a unique prediction of CFR accounts.

3. Experiment 2

There are at least two ways to explain the relationship between causal and counterfactual judgments we observed in Experiment 1. One possibility, argued for by CFR accounts, is that the differences in the relevance of the counterfactual alternatives in the different conditions explains the differences in participants' causal judgments. Specifically, in conjunctive causal structures like the one focused on by Samland and Waldman (2016), the more one focuses on counterfactual alternatives for a specific antecedent event, the more one's attention is drawn to the fact that the outcome depends on that specific event, and thus the more participants should regard that event as the cause of the subsequent outcome (for a more formal treatment, see Icard et al., 2017).

An alternative, and perhaps simpler, way to explain the relationship would be to instead extend Samland and Waldmann (2016)'s approach to explaining the pattern of causal judgments to the pattern of counterfactual relevance judgments. In other words, just as participants may have been interpreting the causal question as being one about moral responsibility, perhaps they were also inclined to interpret the counterfactual relevance question as being one about moral accountability. This alternative is especially plausible given that mental state manipulations in Experiment 1 affected both causal and counterfactual relevance judgments. Naturally, if both questions are simply capturing the same underlying moral accountability judgment, then it should not be surprising that they are highly correlated.

Helpfully, there is a class of cases where the predictions of these two different accounts come apart, namely, cases involving disjunctive causal structures, where multiple antecedent events are each individually sufficient to bring about the outcome. (These are sometimes called cases of overdetermination.) In these cases, CFR accounts predict that the impact of norm violations on causal judgments will be the opposite of the effect observed in Experiment 1.

Specifically, norm violating events that are not necessary for the outcome will be judged to be *less* causal (Kominsky et al., 2015; Icard et al., 2017). At an intuitive level, the reason for this is that when focusing on counterfactual alternatives to some specific antecedent event, one's attention should be drawn to the fact that the outcome does not actually depend on that specific event. In other words, participants should still consider counterfactual alternatives in which a norm-violating antecedent does not occur, but in such counterfactual possibilities the norm-conforming event will continue to occur, and thus the outcome will persist despite the absence of the norm-violating event. Thus, causal judgments in such overdetermined cases are driven by the increased emphasis on the fact that the outcome did not actually depend on the occurrence of the norm violating event (again, see Icard et al., 2017 for a formal treatment).

Returning to the relationship between causal and counterfactual judgments, what is critical for our purposes is that CFR accounts predict that they should be highly positively correlated in conjunctive causal structures (as found in Experiment 1) but not in disjunctive causal structures. In disjunctive causal structures, CFR accounts hold that the relevant counterfactuals will be one in which the norm violation does not occur. More specifically, if the independently-sufficient cause is norm-conforming, then relevant counterfactual would be one in which the norm violation does not occur, but the outcome happens nevertheless because the norm-conforming event does occur and is sufficient for bringing about the outcome. Of course, this will only occur if people perceive the other antecedent event as normative. If it is not perceived as normative, then people should instead consider as relevant counterfactual possibilities in which neither antecedent occurs, and in these possibilities, the outcome will not occur. Accordingly, CFR accounts predict either an negative correlation between counterfactual relevance ratings and causal judgments of the norm-violating event or no consistent relationship

(depending on the judged normality of the alternative antecedent event). In either case though, the key prediction for our purposes holds: CFR accounts predict that the relationship between causal and counterfactual judgments should differ between conjunctive and disjunctive causal structures. In sharp contrast, an account according to which both causal and relevance questions are understood as questions about moral accountability should predict that they will be highly correlated regardless of the causal structure, since they will be treated as variations of the same underlying question.

3.1. Methods.

3.1.1. Participants. 603 participants (M_{age} = 37.03, SD_{age} = 12.04; 271 female, 3 non-binary) from Amazon Mechanical Turk participated and were paid $0.30 in compensation for their time. Participant recruitment was again automated through TurkPrime.

3.1.2. Stimuli and procedure. This experiment comprises a partial replication and extension of Experiment 1 in Icard et al. (2017). As such, it consisted of a 2 (Norm condition: Norm violation vs. No norm violation) × 2 (Causal structure: Conjunctive causal structure vs. Disjunctive causal structure) × 3 (Scenario: Motion detector vs. Battery vs. Train) between-subjects design[4]. Participants were asked to complete both the causal measure used by Icard et al. (2017) and a counterfactual relevance measure similar to that used in Experiment 1. All participants responded to both measures in a counterbalanced order. The study materials were presented in Qualtrics (Qualtrics, 2005).

[4] In a pretest involving these scenarios, we observed that one of the scenarios used in Icard et al. (2017) did not actually elicit the intended pattern of perceived norm violations (in explicit judgments of which agents violated a norm, rather than causal judgments). Specifically, in the Email scenario, participants regarded both agents as having violated a norm, even though the intended manipulation was for only one agent to have done so. This led the majority of participants in this scenario to fail the relevant control questions. Rather than planning to exclude a large number of participants, we simply decided to exclude this scenario from our replication.

We illustrate the design of this study with an example of all four conditions in one of the three scenarios used (Table 1; the other scenarios can be found in the Supplemental Materials). After reading a randomly assigned vignette, participants answered two questions in random order. The causal measure asked them whether they agreed or disagreed that the norm violating agent caused the subsequent outcome. Participants responded on a scale from 0 ('Disagree') to 100 ('Agree'). The counterfactual relevance measure asked them how relevant it was to focus on the norm violating agent when considering how things could have gone differently, they responded on a scale from 0 ('Not at all relevant') to 100 ('Highly relevant').

Subsequently, participants answered two control questions, one of which asked them to indicate which of the agents in the vignette had done something wrong, and the other of which checked their comprehension of the causal structure in the vignette they read. They then completed an open-ended question asking about the factors that influenced their judgments and a brief demographic questionnaire.

----- Insert Table 1 about here ----

3.2. Results.

We excluded participants who did not answer both of the control questions correctly (127/603 or ~21%; Icard et al. (2017) excluded 54/480 or ~11%) and analyzed the remaining 476 participants' judgments. Mean ratings by question and condition are shown in Fig. 2. All reported analyses had power ≥ 93% to detect their observed effects unless otherwise noted.

3.2.1. Replication of Icard et al. (2017). First, we asked whether we replicated the finding that norm violations have inverse effects on causal judgments in conjunctive vs. disjunctive

causal structures. To do this, we analyzed agreement with the causal statement by comparing a series of linear mixed-effects models using the lme4 package in R (Bates, Maechler, Bolker, Walker, et al., 2014). This analysis revealed a main effect of Norm, $\chi^2(1) = 11.73$, $p < .001$, and a main effect of Causal Structure, $\chi^2(2) = 36.01$, $p < .001$. Critically, however, these effects were qualified by a significant Norm × Causal Structure interaction, $\chi^2(1) = 66.60$, $p < .001$, replicating Icard et al. (2017). We next decomposed this interaction by separately analyzing the effect of a norm violation in conjunctive vs. disjunctive scenarios. In conjunctive scenarios, participants *more* agreed with the causal statement when the agent violated a norm ($M = 81.7$; $SD = 22.4$) than when the agent did not ($M = 50.2$; $SD = 32.2$), $\chi^2(1) = 70.56$, $p < .001$. By contrast, in disjunctive scenarios, participants *less* agreed with the causal statement when the agent violated a norm ($M = 40.8$; $SD = 43.4$) than when the agent did not ($M = 56.0$; $SD = 30.2$), $\chi^2(1) = 10.56$, $p = .001$.

3.2.2. Primary analyses. Having replicated the pattern of causal judgments predicted by CFR accounts and observed in Icard et al. (2017), we next wanted to ask whether the pattern of counterfactual relevance judgments differed from the pattern of causal judgments in the disjunctive, but not conjunctive, scenarios, as predicted by CFR accounts. An obvious alternative possibility is that the two judgments would mirror each other across both causal structures, as predicted by accounts according to which both causal and counterfactual questions are interpreted as questions about moral responsibility. Statistically, we can ask which of these two predictions is correct by treating our two DVs as a within-subjects factor of Question (Causal vs. Counterfactual relevance), and testing for a Norm × Causal structure × Question three-way interaction, which would indicate that the Norm × Causal structure interaction observed for causal judgments did not arise for judgments of counterfactual relevance. Indeed, this three-way

interaction effect was highly significant, $\chi^2(1) = 42.95$, $p < .001$. Moreover, this interaction effect was driven by the fact that, while we observed a Norm × Causal structure interaction for causal judgments (see §3.2.1. above), we did not observe one for judgments of counterfactual relevance, $\chi^2(1) = 1.08$, $p = .299$. In contrast, we find that participants found it more relevant to consider counterfactuals focused on the norm violating agent *both* in conjunctive causal structures, $\chi^2(1) = 75.18$, $p < .001$ and in disjunctive causal structures, $\chi^2(1) = 43.65$, $p < .001$, as predicted by CFR accounts (Fig. 2).

----- Insert Figure 2 about here ----

3.2.3. Relationship between causal and counterfactual judgments. To further explore the contrasting predictions of CFR and polysemy accounts, we next asked whether, at the level of both participants and conditions, judgments of counterfactual relevance predicted causal judgments in conjunctive scenarios *but not disjunctive scenarios*. We found that this was indeed the case (Fig. 3). Specifically, in the conjunctive scenarios, we found that counterfactual relevance judgments were tightly correlated with causal judgments at both the item-level, $r = 0.959$, $p < .001$, and the participant-level, $r = 0.612$, $p < .001$, mirroring the pattern found in Experiment 1. In the disjunctive scenarios, by contrast, we found that counterfactual relevance judgments were slightly (but not significantly) *negatively* correlated with causal judgments at the item-level, $r = -0.221$, $p = .675$,[5] but slightly positively correlated at the participant-level, $r = 0.165$, $p = .014$, suggesting no stable relationship.

[5] At the item level we had 80% power to detect a correlation of $r = .72$ or greater, which we expected to be adequate to detect a true effect, since the item-level correlation observed in Experiment 1 was $r = .85$.

----- Insert Figure 3 about here ----

3.3. Discussion.

This study was designed to test two competing explanations of the relationship between causal and counterfactual judgments observed in Experiment 1. The results provide clear evidence against an explanation according to which both the causal and the counterfactual questions were interpreted as being about moral accountability. This explanation predicts that causal and counterfactual judgments will be closely related even in cases involving disjunctive causal structures (after all, they should be interpreted as two versions of the same underlying question). We did not find this to be the case.

Instead, the results provide clear evidence in favor of the unique predictions of CFR accounts. In conjunctive causal structures, because the outcome depends on both antecedent events, the more relevant one finds it to consider counterfactual alternatives to a specific event, the more one should regard that event as the cause of the outcome. In disjunctive causal structures however, this dependence relation does not hold, and thus CFR accounts do not predict a positive correlation between counterfactual relevance and causal judgments. The results from our experiment confirm this difference in the relationship between causal and counterfactual judgments. Critically, as predicted by CFR accounts, counterfactual relevance was still highly sensitive to norm violations, it was simply that because the counterfactuals do not affect causal judgments in the same way in disjunctive causal structures, we no longer see this difference in counterfactual relevance resulting in a corresponding difference in causal judgments.

There are many differences in the methods of Experiment 1 and 2, including the contextual backstory, the use of causal strength judgments rather than causal selection, the

valence of the outcome, and so on. Thus, this experiment clearly does not provide a direct disjunctive 'control' for Experiment 1. Importantly however, the aim of this study was not to investigate anything specific about the particular experimental paradigm used in Experiment 1. Rather, it was to ask a more general question about whether the widely-observed correlation between causal and counterfactual relevance judgments could be explained by *both* questions being interpreted as questions about moral responsibility. Indeed, even though the valence of the outcome was variable between scenarios in this experiment, and different from the outcome in Experiment 1, because the outcome valence was held constant across causal structures, it cannot explain the difference between the conjunctive an disjunctive conditions. The results cast doubt on this explanation and provide evidence in favor of CFR accounts.

While promising, the evidence presented thus far has been primarily correlational and does not yet provide direct support for the key causal claim of CFR accounts. To provide clear empirical evidence for the *mechanism* proposed by CFR accounts, one would need to *directly* manipulate the relevance of counterfactual alternatives (independent of any norm violation) and show that participants' causal judgments are affected in a qualitatively similar way. We explore this next.

4. Experiment 3

Our way of elaborating CFR accounts explains the pattern observed in Samland and Waldmann (2016) by arguing that the changes in the relevance of counterfactual alternatives for the agent (but not the use of the artifact) *caused* participants to see the agent (but not the use of the artifact) as the cause of the outcome. This proposed mechanistic relationship between counterfactual and causal judgments is meant to be perfectly general, i.e., specific to neither

agents nor moral norm violations: One should observe the same pattern of effects on causal judgment if one simply asked participants to explicitly consider relevant ways that a (perfectly norm-conforming) agent could have acted differently (Phillips et al., 2015, cf. Samland & Waldmann, 2016, Experiment 2). Therefore, if one explicitly asks participants to consider alternatives to how the artifact functioned, CFR accounts predict a corresponding increase in causal judgments of the artifact (provided the outcome counterfactually depends on the artifact).

To test these hypotheses, we presented participants with a vignette involving a norm-conforming agent who uses an inanimate artifact which then leads to an outcome. We then directly ask participants to consider relevant counterfactual alternatives either for how the agent acted or the inanimate artifact functioned and then measured the effect of this manipulation on participants' causal judgments of both the agent and the artifact used. We explore this prediction using both a causal selection measure (Experiment 3a) and a causal strength measure (Experiment 3b). We examined measures of both causal selection and causal strength because (i) we are attempting to understand the mechanism behind effects on both kinds of measures (selection in Experiment 1 and strength in Experiment 2) and (ii) CFR accounts predict that this counterfactual manipulation should have roughly the same effect in either case.

4.1. Methods.

4.1.1. Participants. In Experiment 3a, 602 participants (M_{age} = 37.74, SD_{age} = 12.28; 335 female, 2 unreported) and in Experiment 3b, 601 participants (M_{age} = 35.96, SD_{age} = 15.58; 304 female, 2 unreported) were recruited from Amazon Mechanical Turk and were compensated with $0.25 for their time. Participant recruitment was automated through TurkPrime.

4.1.2. Stimuli and procedure. Both Experiment 3a and 3b used a 3 (Counterfactual condition: Agent Counterfactual vs. Artifact Counterfactual vs. No Counterfactual) × 2 (Question: Agent Question vs. Artifact Question) design. Counterfactual condition was manipulated between-subjects in both experiments. In Experiment 3a, Question was a between-subjects factor; and in Experiment 3b, Question was a within-subjects factor. The study materials were presented in Qualtrics (Qualtrics, 2005).

Participants in both experiments read a vignette involving a vending machine in an academic department (see Table 2, left column). The machine had three levers (red, black, and white): The red lever and black lever both produce pencils, and the white lever produces erasers and, due to a malfunction, broken pencils. (The white lever is never used but was included for consistency with Experiment 4; see Table 2.) There were also two agents: an administrative assistant and Professor Smith (a recent hire who did not know about the malfunctioning lever). Both administrators and faculty were allowed to take pencils from the machine. Both the administrative assistant and Prof. Smith request pencils using the black and red levers, which both function appropriately. This results in a problem later when a student who needs a pencil cannot get one, because the machine is out of pencils.

----- Insert Table 2 about here ----

After reading the vignette, participants underwent the counterfactual manipulation. In the Agent Counterfactual condition, for example, participants were asked to think about Professor Smith's decision to take a pencil from the vending machine and then to consider and describe one relevant way that things could have gone differently such that the professor would not have

taken one of the pencils from the vending machine. Note that, in asking about the decision, we explicitly test our elaboration of CFR accounts: That the counterfactual relevance (and thus causal judgments) of the agent's decision to act can be separated from components of the execution of the action itself. In the Artifact Counterfactual condition, by contrast, participants were instead asked to consider and describe a relevant way in which the red lever could have functioned differently such that it did not produce a pencil from the vending machine. In the No Counterfactual condition, participants were simply asked to describe the story they read.

In Experiment 3a, participants completed a causal selection measure which asked them to judge who or what caused the problem, similar to Experiment 1. In the Agent Condition, participants could select either Professor Smith or the Administrative Assistant (or both or neither). In the Artifact Condition, participants could select either the Red Lever or the Black Lever (or both or neither).

In Experiment 3b by contrast, participants instead rated their agreement (on a scale from 0 ['Completely disagree'] to 100 ['Completely agree']) with a statement that the Professor caused the problem, and then separately with a statement that the red lever caused the problem, a measure that is typically interpreted as capturing causal strength (e.g., Icard et al., 2017). These statements were presented in counterbalanced order and on separate pages.

In both experiments, participants then completed a pair of control questions that asked them about which levers were actually pulled and about who actually received a pencil in the original story.

4.2. Results.

4.2.1. Experiment 3a Results. We excluded participants who did not answer both of the check questions correctly (169/602 or ~28%) and analyzed the remaining 433 participants' judgments. All reported analyses had power ≥ 99% to detect their observed effects unless otherwise noted. To facilitate comparison of participants' judgments, we computed a measure of participants' preference for selecting the counterfactual-focus event as a cause. Participants who selected *only* the event that they considered counterfactual alternatives to were assigned a score of 1, participants who selected *both* or *neither* events as causes were assigned score of 0, and participants who selected only the event they did *not* consider counterfactual alternatives to were assigned a score of -1. Mean causal preference scores by question and condition are shown in the left panel of Fig. 4, and the raw frequency of each response is reported in Table 3a and 3b.

We analyzed participants' scores with a 2 (Causal question: Agents vs. Artifacts) × 3 (Counterfactual condition: Professor vs. None vs. Lever) proportional odds logistic regression. This analysis revealed a main effect of Counterfactual condition, (LRT=22.20 [df=2], p<.001), no main effect of Causal question (LRT=0.089 [df=1], p=.925),[6] and critically a Counterfactual condition × Causal question interaction effect (LRT=17.63 [df=2], p<.001).

We decomposed this interaction effect by separately comparing participants' causal preference scores in the two Question conditions. In the agents condition, participants tended to select Professor Smith as the cause of the problem more when they considered alternatives to Professor Smith's action than when they did not consider any counterfactual alternatives (LRT = 11.62 [df = 1], p < .001), or when they considered counterfactual alternatives to what the Red Lever did (LRT = 3.65 [df = 1], p = .056), though the latter difference was only marginally significant. In the Artifacts condition, we observed the mirror-image of this pattern: Participants

[6] This analysis had 80% power to detect an LRT ≥ .14.

tended to select the Red Lever as the cause of the problem more when they considered alternatives to what the Red Lever did than when they did not consider any counterfactual alternatives (LRT = 27.41 [df = 1], $p < .001$), or when they considered counterfactual alternatives to what Professor Smith did (LRT = 12.98 [df = 1], $p < .001$).

----- Insert Table 3a-b about here ----

4.2.2. Experiment 3b Results. We excluded participants who did not answer both check questions correctly (178/603 or ~30%) and analyzed the remaining 423 participants' judgments. All reported ANOVAs were at power ≥ 99% to detect their reported effects. All pairwise comparisons had power ≥ 86% unless otherwise noted. Mean ratings by question and condition are shown in the right panel of Fig. 4. First, we analyzed the agreement with the two causal statements by comparing a series of linear mixed-effects models using the lme4 package in R (Bates et al., 2014). This analysis revealed a main effect of Question, $\chi^2(1) = 53.135$, $p < .001$, and a main effect of Counterfactual condition, $\chi^2(2) = 13.492$, $p = .001$. Critically, however, these main effects were once again qualified by a significant Question × Counterfactual interaction, $\chi^2(2) = 23.04$, $p < .001$. We decomposed this interaction using a series of planned comparisons.

Pairwise comparisons revealed that participants more tended to agree that Professor Smith was a cause of the problem when they considered alternatives to Professor Smith's action (M = 32.99, SD = 33.33) than when they considered alternatives to the way the Red Lever functioned (M = 24.43, SD = 29.12), $t(279) = 2.27$, $p = .024$, $d = 0.272$,[7] or when they did not

[7] This analysis had 62% power for this effect size, and 80% power to detect $d \geq .34$

generate any relevant counterfactual alternatives, (M = 18.22, SD = 27.28), t(282.48) = 4.12, p < .001, d = 0.482.

We also observed a corresponding pattern in participants' agreement with the statement that the Red Lever caused the problem: participants agreed that the Red Lever was more of a cause when they considered alternatives to the way the Red Lever functioned (M = 20.11, SD = 33.34), than when they considered alternatives to what Professor Smith did (M = 10.05, SD = 20.59), t(213.65) = 2.99, p = .003, d = 0.367, or when they did not generate any relevant counterfactual alternatives, (M = 8.62, SD = 19.64), t(211.21) = 3.42, p < .001, d = 0.421.

----- Insert Figure 4 about here ----

4.3. Discussion.

These experiments provide evidence for the key mechanistic claim of CFR accounts by *directly* manipulating the relevance of counterfactual alternatives involving either one of the agents or artifacts, in a context that did not involve any norm violations. In short, we found that manipulating the relevance of counterfactual alternatives to what one of the artifacts did affected causal judgments of that artifact. Similarly, manipulating the relevance of counterfactual alternatives to one of the agent's actions had a corresponding effect on causal judgments of that agent. This was true whether we used an explicit causal selection measure that asked about both agents/artifacts (Experiment 3a) or an agreement rating with a causal claim about the counterfactually-focused agent/artifact (Experiment 3b). These results jointly provide direct support for the causal mechanism suggested by the correlations in Experiment 1 and 2.

It is worth addressing two additional features of these results: First that in Experiment 3b, agreement ratings fell consistently below the midpoint of the agreement rating scale, and second that the effect sizes in both 3a and 3b were relatively small compared to the effect of norm violations in Experiments 1 and 2. With regard to the first point, we suspect that this indicates that many participants were reluctant to assign full causal responsibility to any one particular factor when asked about it in isolation. This interpretation is supported by the fact that in Experiment 3a, which instead used a measure that asked participants to indicate which of the factors were causes, participants overwhelmingly indicated that the event focused on in the counterfactual manipulation was either part of the cause or the sole cause, i.e. they almost never omitted it by selecting the alternative cause alone or choosing neither (Table 3a-b).

As for the effect sizes, while smaller than those in Experiments 1-2, they are in line with previous work that has used explicit counterfactual manipulations (Phillips et al., 2015; Samland & Waldmann, 2016, Experiment 2). A likely explanation for the fact that these effects are generally weaker is that the experimental manipulations used are relatively inefficient ways of getting people to regard counterfactuals as relevant. CFR accounts hold that merely acknowledging the existence of a counterfactual possibility may not affect causal judgment, if the counterfactual is regarded as irrelevant; CFR accounts generally predict that counterfactuals will affect causal judgments only to the extent that they are actually regarded as relevant.

In cases of the norm violations that actually occurred in Experiment 1 and 2, it may have been highly relevant to consider counterfactual alternatives to those events because they intrinsically ‘should not have happened’. In this experiment, by contrast, we provided participants only extrinsic motivation to generate counterfactual alternatives from scratch which they think might be relevant (to some degree), and it is likely that the counterfactuals induced by

this manipulation were not regarded as being as relevant as those induced by cases of genuine norm violations. In other words, participants may be fully aware of the necessity and sufficiency of a given cause in the particular counterfactual they are explicitly asked to consider but are more reluctant to let those counterfactuals inform their causal judgment. Ultimately however, the fact that these manipulations were less effective in increasing the relevance of counterfactual alternatives than genuine norm violations does not undermine the essential claims of CFR accounts, though it may be interesting to investigate in future work.

By confirming that direct counterfactual manipulations affect judgments of inanimate artifacts in the same way that they affect judgments of intentional agents, these results further help explain why Samland and Waldmann (2016) originally observed that moral norm violations differentially affected causal judgments of inanimate artifacts and intentional agents. Moral norm violations simply do not make it more relevant to consider counterfactual alternatives to *the way an artifact was used or the way that artifact functioned*, even when used by agents who were themselves violating norms. Thus, causal judgments of the artifact were unaffected by moral norm violations. If Samland and Waldmann had instead considered cases where the functioning of the artifact violated a norm, they likely would have found an inverse pattern: causal judgments of inanimate artifacts would be more affected than causal judgments of intentional agents. We next turn to testing this specific prediction by investigating cases of prescriptive norm violations that apply specifically to artifacts rather than agents, i.e., norms of *proper function*.

5. Experiment 4

Experiments 1 - 3 address the challenges to CFR accounts raised by Samland and Waldmann (2016) and add novel evidence in support of an elaborated CFR account in scenarios

involving both agents and artifacts. As mentioned above, our elaborated CFR account predicts that if an artifact violates a prescriptive norm of proper function, it should affect causal judgments by changing the relevance of counterfactual alternatives around the behavior of the object in much the same way that moral norm violations affect causal judgments by changing the relevance of counterfactual alternatives around the behavior of an agent. We test this prediction in a final study.

Previous work has mostly focused on moral norm violations and statistical or descriptive norm violations (e.g., Alicke, 2000; Kominsky et al., 2015; Icard et al., 2017). However, there are other types of prescriptive norms which do apply to artifacts, namely, norms of proper functioning. While a few studies have found some evidence that violations of norms of proper functioning affect causal judgments of artifacts (Hitchcock & Knobe, 2009; Livengood et al., 2017), there is currently no direct evidence that these effects are due to changes in the relevance of counterfactual alternatives. Therefore, in Experiment 4, we examined whether prescriptive norm violations that apply to inanimate artifacts affect counterfactual relevance and causal judgments of artifacts in the same way that moral norm violations affect counterfactual relevance and causal judgments of intentional agents.

5.1. Methods.

5.1.1. Participants. 403 participants (M_{age} = 34.96, SD_{age} = 11.90; 205 females, 1 unreported) from Amazon Mechanical Turk participated and were paid $0.25 in compensation for their time. Participant recruitment was again automated through TurkPrime.

5.1.2. Stimuli and procedure. This experiment used a 3 (Norm violation: Immoral vs. Malfunction vs. No violation) × 2 (Question: Agent vs. Artifact) design, administered fully between-subjects. The study materials were presented in Qualtrics (Qualtrics, 2005).

Participants read one of three vignettes involving a vending machine in an academic department as in Experiment 3 (see Table 2). In every condition the machine had three levers (red, black, and white): two that produce pencils and one that produces an eraser but which frequently malfunctioned and also gave a broken pencil. There were also two agents: an administrative assistant, and Professor Smith (a recent hire who did not know about the malfunctioning lever). Prof. Smith always pulls the red lever, and the assistant always pulls the black lever, which later results in a problem for a student who needs a pencil to take a test but cannot get one.

In the No norm violation condition, the red lever and black lever both produce pencils (the malfunctioning white lever plays no role) and both the administrators and the faculty are allowed to take pencils from the machine. Both request pencils using the black and red levers, which both function appropriately. The Immoral condition was identical to the No violation condition, except that the faculty are not allowed to get pencils from the machine (but administrative assistants are allowed to); this rule was known by Prof. Smith. Lastly, the Malfunction condition was identical to the No violation condition except that it was the red (rather than white) lever that produces erasers and consistently malfunctions to also produce a broken pencil. Prof. Smith (who has no way of knowing that the red lever malfunctions) wants an eraser and uses the red lever, which delivers both an eraser and a broken pencil.

Participants were then asked a relevance-of-counterfactual-alternatives question and a causal question in random order on separate pages. The relevance of alternatives question was

worded and presented the same way as Experiment 1, and either focused on the agents (Prof. Smith, administrative assistant) or the artifacts (red lever, black lever). The causal question similarly asked either *who* caused the problem (agent condition) or *what* caused the problem (artifact condition), and participants could select one or both potential causes (similar to the method used by Samland and Waldman [2016], and in Experiments 1 and 3a).

These were followed by three comprehension check questions and two additional manipulation-check questions. The comprehension questions ensured that participants understood the key facts about the levers, agents, and outcome of the scenario. Additionally, participants rated, on a 0-100 scale, how likely the malfunction was to occur, in order to verify that participants did not think the malfunction was a statistical norm violation (the malfunctioning lever was described as very consistently malfunctioning in all conditions). Finally, participants rated their agreement with the statement "It was morally wrong for Prof. Smith to pull the red lever" on a 7-point Likert scale, with the expectation that ratings should be higher in the moral violation condition than the other two conditions, which should not differ from each other.

5.2. Results.

We excluded participants who did not answer all three of the check questions correctly (145/403 or ~36%) and analyzed the remaining 258 participants' judgments. All analyses had power ≥ 99% to detect their reported effects unless otherwise noted. We first analyzed the manipulation-check questions to ensure that we successfully manipulated both the moral status of the Prof. Smith's action and did not inadvertently manipulate whether the lever's malfunctioning was a descriptive norm violation. Both conditions were overwhelmingly met.

5.2.1. Moral check. Participants in the Immoral condition much more strongly agreed that it was immoral for Professor Smith to pull the red lever (M = 3.30, SD = 1.63), than participants in the No norm violation condition (M = 6.22, SD = 1.07), $t(152.97) = -13.92$, $p < .001$, $d = 2.09$, or the Malfunction condition (M = 6.06, SD = 1.34), $t(172) = -12.14$, $p < .001$, $d = 1.84$. The No norm violation and Malfunction conditions did not differ significantly from one another, $t(166) = 0.84$, $p = .40$, $d = 0.04$.[8]

5.2.2. Probability Check. Participants in all three conditions estimated the probability that the lever which gave erasers would malfunction to be well above 50%, and thus was not a descriptive norm violation. Most critically, this was observed in the Malfunction condition (M = 89.33, SD = 13.91), $t(78) = 25.123$, $p < .001$, $d = 2.83$. It was additionally observed in the Immoral condition (M = 75.84, SD = 21.21), $t(85) = 11.30$, $p < .001$, $d = 1.22$, and the No norm violation condition (M = 82.14, SD = 19.07), $t(76) = 14.79$, $p < .001$, $d = 1.69$.

5.2.3. Causal judgments. Results can be found in Table 4a-b. To facilitate comparison of participants' judgments, we computed a measure of participants' preference for selecting the norm-violating event as a cause, similar to the analysis used in Experiment 3a. Participants who selected *only* the norm-violating event as a cause were assigned a score of 1, participants who selected both or neither events as causes were assigned score of 0, and participants who selected *only* the norm-conforming event were assigned a score of -1. These causal preference scores are presented in Fig. 5, left panel. We then analyzed participants' causal preference scores with a 2 (Causal Question: Agent vs. Artifact) × 3 (Norm condition: Immoral vs. Malfunction vs. Normal) proportional odds logistic regression, as in Experiment 3a. This analysis revealed a main effect of Norm condition, (LRT=71.49 [df=2], p<.001), no main effect of Causal question

[8] This analysis had 80% power to detect d ≥ .43.

(*LRT*=0.045 [*df*=1], *p*=.832)[9], and critically a Norm condition × Causal question interaction effect (*LRT*=31.42 [*df*=2], *p*<.001).

We decomposed this interaction effect by separately analyzing participants' causal preference scores for each of the different norm conditions. We first compared the strength of the preference for the norm-violating agent or artifact relative to the no-violation condition in each of the two violation conditions. We found, as expected, that the norm-violating artifact was more strongly preferred in the Malfunction condition than the No Norm Violation condition (*LRT* = 63.31 [*df* = 1], *p* < .001), and similarly that the norm-violating agent was more strongly preferred in the Moral Violation condition than in the No Norm Violation condition (*LRT* = 45.96 [*df* = 1], *p* < .001). However, we also found a small but significant increase in preference for the norm-violating *artifact* in the *Moral* Violation condition, relative to the No Norm Violation condition (*LRT* = 4.48 [*df*=1], *p* = .03). More surprising, we found a strong increase in preference for the norm violating *agent* in the *Malfunction* condition (*LRT* = 20.28 [*df* = 1], *p* < .001).

We therefore conducted further analyses examining whether the effect of each norm violation was *stronger* on the corresponding cause compared to the other cause (i.e., whether the expected effects were stronger than the unexpected effects). When the relevant norm was moral and thus applied to the agent but not the artifact, participants tended to prefer the norm-violating agent as a cause more than they preferred the norm-violating artifact (*LRT* = 15.33 [*df* = 1], *p* < .001). When the relevant norm was functional, and thus the norm applied to the artifact but not the agent, this pattern was reversed: participants tended to prefer the norm-violating artifact as a cause more than the norm-violating agent (*LRT* = 12.36 [*df* = 1], *p* < .001). When there was no norm that applied to either the agent or the artifact, there was small and non-significant

[9] This analysis and the matching null effect below for relevance judgments had 80% power to detect an LRT ≥ 0.17.

preference for the norm-conforming agent but not the artifact (*LRT* = 1.13 [*df* = 1], *p* = .29). We return to the unexpected effects in the General Discussion.

5.2.4. Counterfactual relevance. We next analyzed participants' judgments of the relevance of counterfactual alternatives in the same way. Results can be seen in Table 4c-d and Fig. 5, right panel. Just as with participants' causal judgments, we observed a main effect of Norm condition, (*LRT* = 40.53 [*df* = 2], *p* < .001), no main effect of Relevance question (*LRT* = 0.10 [*df* = 1], *p* = .75), and critically a Norm condition × Relevance question interaction effect (*LRT* = 33.70 [*df* = 2], *p* < .001).

We decomposed this interaction effect by separately analyzing participants' counterfactual preference scores for each of the different norm conditions. Results were similar to what we found for causal judgments. Relevance preference for the norm-violating artifact was higher in the Malfunction condition relative to the No Norm Violation condition (*LRT* = 40.95 [*df = 1*], *p* < .001), and higher for the norm-violating agent in the Moral Violation condition relative to the No Norm Violation condition (*LRT* = 34.82 [*df = 1*], *p* < .001). In contrast to causal judgments, there was no significant preference for the norm-violating artifact in the Moral Violation condition (*LRT* = 1.90 [*df = 1*], *p* = .17), but there was once again a significant preference for the norm-violating agent in the Malfunction condition (*LRT* = 8.20 [*df = 1*], *p* < .001).

We then again compared the size of the expected effects to the size of the unexpected ones. When a moral norm was violated, participants tended to prefer counterfactuals for the agent more than the artifact (*LRT* = 16.63 [*df* = 1], *p* < .001). When the violated norm was functional, this pattern was reversed: participants preferred counterfactuals for the artifact more than the agent (*LRT*=11.20 [*df*=1], *p*<.001). When there was no norm violation that applied to

either the agent or the artifact, there was a small and significant preference for the norm-conforming agent, but not the artifact (LRT=4.48 [df=1], p=.034).

----- Insert Table 4a-d about here ----

----- Insert Figure 5 about here ----

5.2.5. Relationship between causal and counterfactual judgments. Across the variations in the event asked about (Agent vs. Artifact) and Condition (Moral vs. Malfunction vs. No Violation), average counterfactual relevance ratings were highly correlated with average causal ratings, r = .978 [.808, .998], p < .001. Moreover, causal and counterfactual relevance judgments were also correlated at the level of individual judgments, r = . 533 [.440, .615], p < .001 (Figure 6). All data, stimuli, and analysis code are available at: https://osf.io/cp2d5.

----- Insert Figure 6 about here ----

5.3. Discussion.

Causal and counterfactual relevance judgments of inanimate artifacts were affected by violations of norms of proper function in much the same way that judgments of agents were affected by violations of moral norms. On the one hand, our results provide a conceptual replication of the findings in Samland and Waldmann (2016) by demonstrating that when th agent violated a moral norm, judgments of the agent were affected more than the artifact used. On the other hand, we find a complimentary effect for functional artifacts: When the artifact violated a functional norm, judgments of the artifact were affected more than those of the agent

who used it. Most importantly, participants' causal and counterfactual relevance judgments were correlated at the level of both participants and conditions. Taken together, these patterns support our elaborated CFR account and provide evidence against the claim that causal judgments of intentional agents and inanimate artifacts are governed by unrelated underlying cognitive mechanisms.

At the same time, we did observe an unexpected effect of the artifact *malfunction* on judgments of the *agent*, with a weaker (for causal judgments) or nonexistent (for counterfactual judgments) corresponding effect of the *moral* violation on judgments of the *artifact*. This asymmetry is a novel result, and an unexpected one. While one might be tempted to explain this as resulting from participants thinking of the professor as having violated a moral norm by using the malfunctioning lever, we are explicit that the professor is ignorant of the fact that the lever malfunctions and could not have known; moreover, participants' moral judgments of the professor's action in the malfunction condition suggest that they did not regard the action as wrong. We consider other, more promising ways to explain this effect in the General Discussion.

6. General Discussion

Across four experiments, we find robust support for unified counterfactual relevance (CFR) accounts of the impact of norm violations on causal reasoning. Experiment 1 demonstrated that patterns previously interpreted as contradicting CFR accounts are, in fact, perfectly compatible with natural ways of making existing CFR accounts more precise. Experiment 2 then showed that it cannot be the case that both causal and counterfactual relevance questions are interpreted as questions of moral accountability, as these judgments dissociate when CFR accounts predict they should. Experiment 3 validated the proposed causal

directionality of the correlations in Experiments 1 and 2 by showing that causal judgments of both agents and artifacts are affected by direct manipulations of the availability of relevant counterfactuals. Finally, Experiment 4 showed that prescriptive norm violations do affect causal judgments of inanimate artifacts when the norm being violated applies to artifacts, and again replicated the correlation between causal and counterfactual relevance judgments.

6.1. Norm violations affect counterfactual relevance.

The extant literature on causal judgment now provides evidence for three distinct types of norms that all show similar effects: descriptive statistical norm violations (e.g., Kominsky et al., 2015), prescriptive moral norm violations (e.g. Alicke, 2000), and prescriptive functional norm violations (demonstrated here; see also Hitchcock & Knobe, 2009). The demonstration of the extent to which different norms have a similar impact on causal and counterfactual judgments makes a unified, parsimonious explanation increasingly desirable.

It would be challenging to explain the results of these experiments with non-CFR accounts. For moral violations in particular, the primary alternative accounts are those that argue that 'causal' judgments are being interpreted as 'moral responsibility' judgments (Samland & Waldmann, 2016), or argue that they result from motivated reasoning (Alicke et al., 2011). However, such accounts have no obvious way of explaining (i) why causal judgments and counterfactual relevance judgments are affected in the same way by a norm violation when the norm violation is necessary to the outcome, but dissociate when the outcome is overdetermined, (ii) why direct manipulations of counterfactual relevance in the absence of any norm violation affects causal judgments in the same way that norm violations do, or (iii) why violations of *norms of proper functioning* produce the same effect on causal and counterfactual reasoning

about inanimate objects as violations of moral norms do on causal and counterfactual reasoning about intentional agents.

Collectively, the evidence across our four experiments demonstrates that norm violations affect the relevance of counterfactual alternatives, and the relevance of counterfactual alternatives affects causal judgments. This relationship holds across variations in the nature of the candidate cause and variations in the nature of the norm violation, suggesting that these effects arise from general features of causal reasoning, rather than some domain-specific way of reasoning about intentional agents, morality, or the intended meaning of the word 'cause'.

6.2. How are counterfactuals generated?

Our elaborated CFR account makes one critical addition to previous accounts: the idea that a norm violation affects the counterfactual relevance of a fairly restricted event representation (e.g., an agent's decision to violate a norm), but can have a more minimal impact on the counterfactual relevance of other aspects of that event (e.g., the mechanics of how the action was executed). However, our understanding of the underlying mechanism is far from complete. For example, in Experiment 4, we found that when an inanimate artifact malfunctions, causal and counterfactual ratings of the agent who used that artifact increase somewhat, but at the same time, norm violations by the agent who uses the object do not (or only barely) increase causal or counterfactual ratings of the artifact used by that agent. In the rest of this discussion, we propose a (speculative) account of how counterfactual possibilities are generated and considered when making causal judgments and identify some critical questions that will need to be answered in order to test it.

As noted in the introduction, one shortcoming of existing CFR accounts is that they are not specific about the process by which relevant counterfactual possibilities are generated. All versions of the CFR account hold that, in some way, people are constructing a 'normalized' counterfactual version of the event (in which the norm violations are replaced by norm-conforming actions) and then determining the truth value of the sufficiency and necessity conditionals in those counterfactual alternatives (Hitchcock & Knobe, 2009, p. 589). However, what the 'more normal' version of an event actually consists of is left to intuition. Therefore, to explain these patterns with any form of CFR account, it would be helpful to at least point toward a more fleshed-out account of how people may generate 'normalized' counterfactual possibilities.

We sketch an explanation built around a single novel idea: We propose that when people generate counterfactual possibilities, the possibilities they construct often completely exclude some variables in the original scenario. In such a case, it will clearly be impossible to evaluate the necessity or sufficiency of an event which is absent from generated counterfactuals, and thus we would not expect these counterfactuals to affect participants' causal judgments.

In one sense, this proposal departs substantively from previous CFR accounts: It denies a traditional assumption of work that has sought to align ordinary counterfactual reasoning with the formal notion of counterfactual intervention (Pearl, 2000) or the formal notion of counterfactual assessments in possible worlds semantics (Lewis, 1973; Stalnaker, 1968).

In another sense, however, the proposal relies on a relatively obvious and uncontroversial observation. When participants have been asked to report the 'normalized' counterfactual alternatives they generated in prior studies (e.g., Kahneman & Tversky, 1982), their descriptions typically do not mention variables that are the *immediate downstream* consequences of the

original norm violating action. For example, recall the case in which Mr. Jones drives home by an unusual route and on that route goes through a particular intersection where he is hit by a driver under the influence of drugs. Participants' description of the counterfactual alternatives they generated about these events often involve Mr. Jones not taking the unusual route home (the norm violating event), but they do not include descriptions of Mr. Jones 'not going through that intersection' (the immediate downstream consequence of the norm violating event)– this latter variable seems to be *unrepresented* rather than being represented but having a different value. In particular, whereas previous theories might suggest that one can still evaluate the necessity of going through the intersection in this counterfactual possibility, we propose that such an evaluation is impossible. The key difference is that we are suggesting that this event, Mr. Jones going through that intersection, is not being falsified in this counterfactual possibility, it is being dropped. So, the truth value of the necessity condition, "If this event had not occurred, then the outcome would not have occurred", cannot be evaluated in this counterfactual possibility. This idea is the core of our proposal, and critically for our purposes, this relatively small change can help capture both the fact that a moral violation does not affect judgments of a tool being used, and the fact that a malfunctioning tool can affect judgments of the person using it.

6.2.1. Changing versus removing causal variables and the effect of moral norm violations on causal judgments of actions and inanimate objects. Consider the standard norm violation condition of the Samland and Waldmann (2016) vignette used in Experiment 1 as an example. Let us assume that the initial causal model includes Benni's decision to defy Tom's instructions, Benni's actions in fertilizing the plants, and the application of fertilizer Y to the plants as separate causal variables, along with the corresponding variables for Alex, and the outcome

variable of the plants drying up.[10] These correspond to the foci of the different question conditions in Experiment 1. For example, when we ask for a causal judgment of 'Benni', our findings indicate that it is interpreted as a judgment of "Benni's decision to defy Tom's instructions". The 'normal' counterfactual alternative in this case is that Benni follows Tom's instructions and does not use fertilizer Y (or so we intuit, at least). With this model in hand, one approach would be to say that this counterfactual possibility consists of a model in which the variable "Benni defies Tom's instructions" is false, the variable "Benni's actions" (as they occurred in the actual event) is false, the variable "the application of fertilizer Y to the plants" is false, and the outcome is false. Under this view, the necessity of "the application of fertilizer Y" is validated in this relevant counterfactual possibility as both it and the outcome are false, and therefore causal judgments of that event should be increased.

In contrast, on the alternative proposal we've outlined above, another possibility is that the variable representing "the application of fertilizer Y" is simply not included in the counterfactual model that participants generate. Rather than being set to false, let us say that the process of generating this counterfactual, falsifying "Benni defies Tom's instructions", means that fertilizer Y is simply not represented because it is no longer relevant – there are other downstream consequences of this alternative decision that are represented instead. Thus, when participants are asked to make a causal judgment of "the application of fertilizer Y", the relevant counterfactual offers no information about the necessity or sufficiency of this causal variable, because it is simply absent. Therefore, the relevance of this counterfactual should not influence

[10] There are no commonly agreed-upon guidelines for when or how to carve the totality of the things that occurred into distinct variables that are represented as part of the causal graph (Halpern & Hitchcock, 2014), and similarly little empirical work on how such variables are selected to populate the model in the first place (cf. Halpern & Hitchcock, 2010; Goodman, Mansinghka, & Tenenbaum, 2007). In the current state of the literature we would be equally justified in treating all of the causes as a single variable (e.g., Waldmann & Mayrhofer, 2016), as there is no principled account of how events are parsed when creating these causal models.

judgments of "the application of fertilizer Y", and furthermore the counterfactual relevance judgments of "the application of fertilizer Y" should be similarly unaffected (as it is not part of this, or any, relevant counterfactual). However, judgments of Benni (which we claim are interpreted as judgments of "Benni deciding to defy Tom's instructions") are still affected, as this causal variable is still present in the counterfactual possibility, and falsified.

This speculative proposal explains more completely why distal norm violations may have little to no effect on judgments of proximal causes across all of our experiments. However, we also found that a malfunctioning object *does* affect judgments of the agent using it in Experiment 4. This can be accounted for by our sketch of a proposal as well: Previous work has found that people have different intuitive causal structures for physical and psychological events (Strickland, Silver, & Keil, 2017), such that physical events are expected to be deterministic (with a single antecedent cause). When a malfunction occurs, if the causal structure of the mechanism is intuitively thought of as deterministic, then in order to consider a counterfactual possibility in which the malfunction does not occur, people may have to consider one in which its antecedent is present, but altered (Mandel, 2003). Thus, in Experiment 4, people may *not* consider the possibility that Prof. Smith pulls the red lever and the red lever produces only an eraser but not a broken pencil (i.e., altering only the norm violation without changing its antecedent). Instead, they may consider possibilities in which Prof. Smith does not pull the red lever, falsifying both the causal variables "Prof. Smith pulling the red lever" and "the red lever malfunctioning", and leading to increased causal and counterfactual relevance judgments for both. (Note that this kind of change bears some similarity to a "backtracking" counterfactual inference; see Gerstenberg, Bechlivanidis, & Lagnado, 2013; Lewis, 1979; and Rips, 2010.)

While the present sketch of a proposal points to a way of explaining these results, there are many aspects that remain to be filled in and such an account obviously requires further empirical testing. First and foremost there are the open questions of how we parse events when creating causal models and how we associate causal questions with particular elements of those models. When we ask for causal judgments of "Benni", do we mean "Benni's decision", "Benni's actions in general", or "the specific token instances of how this particular action played out"? Second is the question of when causal variables are removed when generating counterfactual alternatives. An independent way of determining whether an event is *represented* in a causal model is needed to test the novel proposal of this speculative account, but it may be difficult to do so without the question itself changing what is in the model (see Goodman et al., 2007 for one interesting approach). These and other challenges will need to be addressed for the CFR account to be a comprehensive explanation of these effects. Nonetheless, while these questions remain to be answered, we have little doubt that CFR accounts in general provide a more complete explanation of these effects than any alternative we have encountered thus far.

7. Conclusion

While we have provided substantial evidence that the effect of norm violations on causal judgments is best explained by some form of counterfactual relevance account, we have also highlighted many of the limitations of current CFR accounts. We regard the CFR account as a strong foundation for a more general account of causal reasoning, but at present it is only a foundation. Future work must aim to build a complete structure atop it.

Acknowledgements

Frist and foremost, we would like to thank Fiery Cushman, for his generous support in this research, and Harvard's Moral Psychology Research lab for feedback on this work. JFK was supported by NIH grant F32HD089595. JSP was supported by grant N00014-19-1-2025 from the Office of Naval Research and grant 61061 from the John Templeton Foundation.

References

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin, 126*, 556–574. http://doi.org/10.1037/0033-2909.126.4.556.

Alicke, M. D., Rose, D., & Bloom, D. (2011). Causation, norm violation, and culpable control. *Journal of Philosophy*, 108(12), 670-696. doi:10.5840/jphil20111081238

Bates, D., Maechler, M., Bolker, B., Walker, S., et al. (2014). *lme4: Linear mixed-effects models using 'Eigen' and S4*. [R package] Retrieved from https://cran.r-project.org/web/packages/lme4/index.html on April 19, 2018.

Bicchieri, C., Muldoon, R., & Sontuoso, A. (2018). Social Norms, in Edward N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy* (Winter 2018 Edition), https://plato.stanford.edu/archives/win2018/entries/social-norms.

Danks, D., Rose, D., & Machery, E. (2014). Demoralizing causation. *Philosophical Studies, 171*(2), 251-277. doi:10.1007/s11098-013-0266-8

Gerstenberg, T., Goodman, N.D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*.

Gerstenberg, T., Bechlivanidis, C., & Lagnado, D. A. (2013). Back on track: Backtracking in counterfactual reasoning. *Proceedings of the 35th Annual Meeting of the Cognitive Science Society.*

Goodman, N. D., Mansinghka, V. K., & Tenenbaum, J. B. (2007). Learning Grounded Causal Models. *Proceedings of the 29th Annual Meeting of the Cognitive Science Society.*

Halpern, J. Y., & Hitchcock, C. (2010). Actual causation and the art of modeling. In R. Dechter, H. Geffner, & J. Y. Halpern (Eds.), *Heuristics, probability, and causality; a tribute to Judea Pearl* (pp. 383-406). College Publications.

Halpern, J. Y., & Hitchcock, C. (2014). Graded causation and defaults. *The British Journal for the Philosophy of Science*, *66*(2), 413–457. http://doi.org/10.1093/bjps/axt050

Hitchcock, C., & Knobe, J. (2009). Cause and norm. *Journal of Philosophy*, *106*(11), 587–612. doi: 10.5840/jphil20091061128

Hume, D. (1748). *An Enquiry concerning Human Understanding*.

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80-93. doi:10.1016/j.cognition.2017.01.010

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological review*, *93*(2), 136-153. doi:10.1037/0033-295X.93.2.136

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2015). Causal superseding. *Cognition*, *137*, 196–209. http://doi.org/10.1016/j.cognition.2015.01.013

Lewis, D. (1973). Causation. *Journal of Philosophy*, 70: 556–67.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs*, *4*(13), 455-476.

Lewis, D. (1973). *Counterfactuals*. Oxford, Oxford University Press.

Livengood, J., Sytsma, J., & Rose, D. (2017). Following the FAD: Folk attributions and theories of actual causation. *Review of Philosophy and Psychology*, 8(2), 273-294. doi: 10.1007/s13164-016-0316-1

Mandel, D. R. (2003). Effect of counterfactual and factual thinking on causal judgements. *Thinking & Reasoning*, *9*(3), 245-265. doi:10.1080/13546780343000231

Morris, A., Phillips, J. S., Gerstenberg, T., & Cushman, F. A. (2019, February 14). Quantitative causal selection patterns in token causation. https://doi.org/10.31234/osf.io/upv8t

Morris, A., Phillips, J. S., Icard, T., Knobe, J., Gerstenberg, T., & Cushman, F. A. (2018, April 26). Causal judgments approximate the effectiveness of future interventions. https://doi.org/10.31234/osf.io/nq53z

Pearl, J. (2000). *Causality : Models, reasoning, and inference*. Cambridge, U.K.; New York: Cambridge University Press.

Phillips, J., Luguri, J. B., & Knobe, J. (2015). Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition, 145*, 30–42. doi:10.1016/j.cognition.2015.08.001

Phillips, J., & Knobe, J. (2018). The psychological representation of modality. *Mind & Language, 33*(1), 65-94. doi:10.3758/bf03209355

Prentice, D. A. (2007). Norms, prescriptive and descriptive. In R. Baumesiter and K. Vohs (Eds.) *Encyclopedia of Social Psychology*, 630-631.

Rips, L. J. (2010). Two causal theories of counterfactual conditionals. *Cognitive Science, 34*(2), 175-221.

Qualtrics. (2005). [Computer Software]. Provo, UT: Qualtrics.

Samland, J., Josephs, M., Waldmann, M. R., & Rakoczy, H. (2015). The role of prescriptive norms and knowledge in children's and adults' causal selection. *Journal of Experimental Psychology. General*, *145*(January), 125–130. doi:10.1037/xge0000138

Samland, J., & Waldmann, M. R. (2016). How prescriptive norms influence causal inferences. *Cognition*, *156*, 164–176. doi:10.1016/j.cognition.2016.07.007

Stalnaker, R. C. (1968). A theory of conditionals. In *Ifs* (pp. 41-55). Springer, Dordrecht.

Strickland, B., Silver, I., & Keil, F. C. (2017). The texture of causal construals: Domain-specific biases shape causal inferences from discourse. *Memory & Cognition*, *45*(3), 442-455. doi:10.3758/s13421-016-0668-x

Sytsma, J., Livengood, J., & Rose, D. (2012). Two types of typicality: rethinking the role of statistical typicality in ordinary causal attributions. *Studies in History and Philosophy of Biology and Biomedical Sciences*, *43*(4), 814-820. doi:10.1016/j.shpsc.2012.05.009

Waldmann, M. R., & Mayrhofer, R. (2016). Chapter Three - Hybrid Causal Representations. *Psychology of Learning and Motivation*, *65*, 85-127. doi:10.1016/bs.plm.2016.04.001

Tables

1a) Norm violation: Suzy and Billy are working on a project that is very important for our nation's security. The boss tells them both: 'Be sure that you are here at exactly 9am. It is absolutely essential that you arrive at that time.'	**1b) No norm violation:** Suzy and Billy are working on a project that is very important for our nation's security. The boss tells Suzy: 'Be sure that you are here at exactly 9am. It is absolutely essential that you arrive at that time.' Then he tells Billy: 'Be sure that you do not come in at all tomorrow morning. It is absolutely essential that you not appear at that time.'
2) Event: Both Billy and Suzy arrive at 9am.	
3a) Conjunctive structure: As it happens, there was a motion detector installed in the room where they arrived. The motion detector was set up to be triggered if *more than one person* appeared in the room at the same time. So the motion detector went off.	**3b) Disjunctive structure:** As it happens, there was a motion detector installed in the room where they arrived. The motion detector was set up to be triggered if *at least one person* appeared in the room. So the motion detector went off.
Causal Measure: How much do you agree with the following statement? Billy caused the motion detector to go off.	
Counterfactual Relevance Measure: Now suppose that some people are discussing this story and wondering how things could have been different. In thinking about who could have acted differently, please tell us how relevant or irrelevant it would be to focus on the following: Billy	

Table 1. Example vignette from Experiment 2 (Motion detector) illustrating both norm violation and the causal structure information.

<table>
<tr><th>No violation (Exp. 3 &4)</th><th>Moral violation (Exp. 4)</th><th>Malfunction (Exp. 4)</th></tr>
<tr><td colspan="3">A philosophy department at a small university bought a vending machine to dispense and keep track of office supplies.</td></tr>
<tr><td colspan="2">Currently, the machine has pencils, which you can get by pulling a black lever or a red lever, and erasers, which you can get from the machine by pulling a white lever.

The white lever on the vending machine frequently malfunctions. When pulled, it produces both an eraser and a pencil, but breaks the pencil in the process.</td><td>Currently, the machine has pencils, which you can get by pulling a black lever or a white lever, and erasers, which you can get from the machine by pulling a red lever.

The red lever on the vending machine almost always malfunctions. When pulled, it produces both an eraser and a pencil, but breaks the pencil in the process.</td></tr>
<tr><td>Pencils are often needed by students who are taking tests, but there are usually plenty of pencils and erasers in the vending machine, so both the administrative assistants and the professors are allowed to take pencils and erasers from the machine.
Professor Smith was recently hired at the department. He was told that he is allowed to take pencils and erasers from the vending machine. However, he was not told that the white lever almost always malfunctions.</td><td>Pencils are often needed by students who are taking tests, so to make sure there are always enough pencils, the philosophy professors are told to buy their own pencils and not take them from the vending machine. Only the administrative assistants are allowed to take pencils from the vending machine. There are plenty of erasers though, so both administrative assistants and professors are allowed to take erasers from the machine.
Professor Smith was recently hired at the department. He was told that he was not allowed to take pencils from the vending machine, but that he was allowed to take erasers. However, he was not told that the white lever almost always malfunctions.</td><td>Pencils are often needed by students who are taking tests, but there are usually plenty of pencils and erasers in the vending machine, so both the administrative assistants and the professors are allowed to take pencils and erasers from the machine.
Professor Smith was recently hired at the department. He was told that he is allowed to take pencils and erasers from the vending machine. However, he was not told that the red lever almost always malfunctions.</td></tr>
<tr><td colspan="3">One morning, Professor Smith and an administrative assistant both go to the vending machine and pull a lever at the same time. The administrative assistant wanted a pencil, so he pulled the black lever.</td></tr>
<tr><td colspan="2">Professor Smith also wanted a pencil and so he pulled the red lever. Both levers worked like they were supposed to, and they both got a pencil from the vending machine.</td><td>Professor Smith wanted an eraser and so he pulled the red lever. The black lever worked like it was supposed to and the administrative assistant got a pencil. However, the red lever malfunctioned, and the professor got an eraser and a completely broken pencil.</td></tr>
<tr><td colspan="3">These were the last two pencils in the vending machine. Not long after, a student desperately needs a pencil for a test, but when she goes to the vending machine, there are no more pencils. This leads to a serious problem.</td></tr>
</table>

Table 2. Vignettes used in Experiments 3 and 4. The “no violation” version (left column) was used in both experiments, the moral violation (center column) and malfunction (right column) versions were only used in Experiment 4. Cells that overlap multiple columns contain text that was identical between those conditions. NOTE: Bold text is used here highlight differences between conditions but was not included in the text participants read.

Table 3a.	**Agents condition selections**			
CF Condition:	Admin only	Neither	Both	Prof only
Professor	7	0	35	**44**
Red Lever	5	1	40	23
None	7	0	53	17

Table 3b.	**Artifacts condition selections**			
CF Condition:	Black lever only	Neither	Both	Red lever only
Professor	9	10	27	23
Red Lever	2	5	15	**38**
None	9	10	39	14

Table 3a-b. Counts of participants' causal selection responses for each of the different Question conditions in Experiment 3a (Table 3a and 3b), and for each of the different Counterfactual conditions (rows). The responses for the target of the counterfactual manipulation are highlighted in bold.

Table 4a.	**Agents condition causal judgments**			
Norm condition:	Admin only	Neither	Both	Prof only
Immoral	1	0	7	**36**
Malfunction	6	0	9	30
No violation	8	0	26	4

Table 4b.	**Artifacts condition causal judgments**			
Norm condition:	Black lever only	Neither	Both	Red lever only
Immoral	7	1	18	20
Malfunction	0	0	2	**38**
No violation	6	8	24	7

Table 4c.	**Agents condition counterfactual relevance judgments**			
Norm condition:	Admin only	Neither	Both	Prof only
Immoral	2	2	12	**28**
Malfunction	6	10	13	16
No violation	6	10	21	1

Table 4d.	**Artifacts condition counterfactual relevance judgments**			
Norm condition:	Black lever only	Neither	Both	Red lever only
Immoral	3	18	16	9
Malfunction	1	1	10	**28**
No violation	1	24	18	2

Table 4a-d. Counts of participants' responses for each of the different conditions in Experiment 4 (Causal judgments Table 4a-b, relevance judgments Table 4c-d) and for each of the different Norm violation conditions (rows). The responses for the norm violating entity are highlighted in bold.

Figure captions

Figure 1. Depiction of the relationship between participants' causal judgments in Experiment 1 (y-axis) and the previous causal results from Samland and Waldmann (2016) (left panel) and the mean relevance judgments in the current experiment (right panel). Judgments related to the norm-conforming agent, action, or use of artifacts are marked with a 'C'; Judgments related to the norm-violating agent are marked with a 'V'. The norm labels use Samland & Waldmann's behavior-based understanding of norms to facilitate comparison.

Figure 2. Average agreement ratings with the causal statement (left graph) and counterfactual relevance statement (right graph) as a function of whether the agent violated a norm (red bars) or did not (blue bars) both in conjunctive causal structures (left panels) and disjunctive causal structures (right panels). Error bars depict +/- 1 SEM.

Figure 3. Relationship between causal judgments and judgments of the relevance of counterfactual alternatives. Solid lines represent the linear relationship between participant-level pairs of responses (depicted by the smaller points). Dotted lines represent the linear relationship between the item-level causal and counterfactual relevance judgments (depicted by the larger points). The color of each points indicates whether the judgment was made when the agent violated a norm (red points), no norm was violated (blue points). The shape of each point indicates the scenario in which the judgment was made. Error bars depict +/- 1 SEM.

Figure 4. Average preference score (see § 4.2.1) for the causal selection measure in Experiment 3a (left graph) and agreement with the causal statement in Experiment 3b (right graph) for the Agent (left panels) and Artifact (right panels) as a function of whether the counterfactual

manipulation focused on the Professor (red bars), there was no counterfactual manipulation (blue bars), or the counterfactual manipulation focused on the Red Lever (yellow bars). Error bars depict +/- 1 SEM.

Figure 5. Average preference score for the norm-violating event in causal judgments (left plots) and counterfactual relevance judgments (right plots), as a function of whether the questions focused on intentional agents (left panels) or functional artifacts (right panels). The color of the bars indicates whether the agent violated a norm (red bars), no norm was violated (blue bars), or the functional artifact violated a norm (yellow bars). Error bars depict +/- 1 SEM.

Figure 6. Relationship between causal judgments and judgments of the relevance of counterfactual alternatives. The solid line represents the linear relationship between participant-level pairs of responses (depicted by the smaller points). The dotted line represents the linear relationship between the condition-level causal and counterfactual relevance judgments (depicted by the larger points). The color of each points indicates whether the judgment was made when the intentional agent violated a norm (red points), no norm was violated (blue points), or the functional artifact violated a norm (yellow points). The shape of each point indicates whether the judgments were made of functional artifacts (circles) or intentional agents (triangles). Error bars depict +/- 1 SEM.