Immoral professors and malfunctioning tools:

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

Jonathan F. Kominsky*†

Harvard University, Department of Psychology

Jonathan Phillips*

Harvard University, Department of Psychology

*Authors contributed equally to this work

†Corresponding author

Address for reprints and correspondence	:	Jonathan F. Kominsky William James Hall #1154 Department of Psychology, Harvard University 33 Kirkland St. Cambridge, MA 02138
Email	:	jkominsky@g.harvard.edu
Phone/Fax	:	(617)-384-7930

Word count: 11,016

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 the 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 inanimate artifacts, and therefore contradicting a unified counterfactual analysis of causal reasoning about agents and artifacts. Four experiments provide evidence against this conclusion. Experiment 1 demonstrates that these newly observed patterns in causal judgments are closely correlated with judgments of counterfactual relevance. Experiment 2 then more directly manipulated the relevance of counterfactual alternatives and finds that causal judgments of intentional agents and inanimate artifacts are similarly affected. Finally, Experiments 3 and 4 show that, across variations in causal structure, 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

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). Specifically, people are more inclined to judge that *e* was the cause of *o* if *e* was either very unlikely to happen or morally prohibited. Despite the wide-spread agreement on the existence of the phenomenon, there has been little corresponding agreement on 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; 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 the one hand, researchers have argued that the impact of both statistical and moral norms is best explained by changes in the relevance of counterfactual possibilities. These counterfactual relevance accounts (CFR accounts hereafter) propose that when a norm violation occurs, it increases the relevance of counterfactual alternatives wherein the norm violations are replaced by norm-conforming events (e.g., Halpern & Hitchcock, 2015; Icard et al., 2017). In support of this account, recent work demonstrated that norm violations affect explicit assessments of counterfactual relevance in precisely the same way that they affect causal judgments (Phillips, Luguri, & Knobe, 2015).

On the other hand, other researchers have argued that the impact of violations moral norms is not the same as other kinds of norm violations. One recent approach 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 (Alicke et al., 2011; Samland & Waldmann, 2016; Sytsma, Livengood, & Rose, 2012). On this view, the effect of statistical norms is taken to arise from the ordinary processes involved in counterfactual cognition. In contrast, the impact of violations of moral norms is instead accounted for 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, and thus they report that agents who violate moral norms are more causal, perhaps as a way of assigning blame to them for negative outcomes (Alicke et al., 2011).

In support of this latter view, Samland and Waldmann (2016) (S&W hereafter) reported two important new data points: First, the violation of moral norms selectively influences causal judgments about whether *agents* caused an outcome, but not causal judgments of whether *inanimate artifacts* used by those agents 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 argue that this conclusion is not warranted in three independent ways: (1) We demonstrate that the pattern of causal judgments that was meant to provide evidence against CFR accounts is completely in line with them: participants' judgments of counterfactual

relevance perfectly predict their causal judgments in these cases. (2) We provide explicit evidence that causal judgments of both agents and artifacts are affected by direct manipulations of counterfactual relevance. (3) We show that prescriptive norms violations do affect causal judgments of inanimate artifacts when the norm being violated actually applies to artifacts (i.e., a norm of proper functioning), and again replicate the tight relationship between causal and counterfactual relevance judgments in such cases.

1.1. What is the moral thing for an artifact to do?

In demonstrating that a moral violation affects causal judgments of the agent but not the artifact that the agent uses, S&W argue that CFR accounts are committed to predicting a different pattern. In particular, they suggest that the moral violation should lead people to consider counterfactual alternatives involving differences in both what the agent and in what the artifact did, thereby increasing causal judgments of both. Based on their results (that only causal judgments of the agent are affected), they conclude that these cases represent an instance in which causal judgments are dissociated from the relevant counterfactuals. However, they do not empirically examine the counterfactuals participants regarded as relevant in these cases, and so the putative dissociation hinges on the prediction they ascribe to CFR accounts.

The ascribed prediction misses an essential piece of how CFR accounts are typically framed. 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). Put differently, deviations from 'normality' that are necessary for a given outcome to occur lead to increased

causal attribution to the abnormal event. 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 whether this change in the agent's actions alters the occurrence of the outcome in these counterfactual possibilities; if so, then we would expect the agent to tend to be seen as a clear cause of the outcome.

Moral prescriptions, however, only apply to moral agents. What is the moral thing for an inanimate artifact to do? While there may be morally prescribed actions for agents who use artifacts, there are no morally prescribed actions for inanimate artifacts themselves. It is not a *moral* violation for a wheel to roll down a hill, even if there is a sign expressly forbidding such behavior. Thus, we would not expect people to consider counterfactual alternatives to what an artifact does on the basis of a *moral* norm being violated.

We are therefore left with an unresolved question that needs to be addressed: Does the moral violation by the agent lead participants to consider counterfactual alternatives to both what the agent and the artifact does, or only the agent? We tested this in Experiment 1 with S&W's scenario but using judgments of counterfactual relevance: "how relevant is it to consider how the behavior of [the norm-violating agent/the artifact they used] could have been different?" Unlike causal judgments, counterfactual relevance judgments do not admit of any kind of potential ambiguity whereby they could be reinterpreted as questions of moral responsibility. So, S&W and related accounts should predict that, in the case of a moral norm violation, both agent and artifact should be seen as relevant or neither should be seen as relevant: a pattern that demonstrates counterfactual judgments dissociate from the pattern of causal judgments they previously observed. In contrast, CFR accounts predict that counterfactual relevance judgments

should correlate closely with causal judgments across conditions, and thus that counterfactual relevance judgments should show an increase for the norm-violating agent but not the artifact they use.

This pattern, if confirmed, raises a further question: Why are counterfactual and causal judgments of inanimate artifacts not affected by the norm violations of the agents using them? One possibility is that causal judgments of inanimate artifacts used as tools are simply not affected by counterfactual reasoning. An alternative possibility is that the specific counterfactual alternatives made relevant by moral norm violations are unlikely to affect causal judgments of inanimate objects. For example, if people consider counterfactual alternatives that focus on changing the agent's behavior, there may be no need to include the particular artifact at all; people may entertain counterfactuals in which the agent does something completely unrelated to the artifact. In such cases, the inanimate artifact may simply not be represented or reasoned over, and consequently causal judgments of the inanimate object should be unaffected.

This latter possibility predicts that neither counterfactual relevance judgments nor causal judgments of the artifact should be affected by moral violations (and these judgments should be closely correlated). However, to further show that judgments of inanimate artifacts can be affected by the consideration of counterfactual alternatives when those alternatives actually involve changing the behavior of the artifact, we need a contrast case in which we specifically make counterfactuals to what the *artifact* does relevant, in which case we predict causal judgments of the *artifact* should be increased.

To test this hypothesis in its most general form, we can temporarily put norm violations aside. Regardless of how a counterfactual alternative is made more relevant, it should affect causal judgments. So, in general, if participants are asked to consider relevant alternatives to

what the artifact did, it should affect causal judgments of the artifact. If instead they are asked to consider relevant counterfactual alternatives to what the agent did (allowing them to ignore the artifact), judgments of the artifact are likely to be less affected. Even if we present participants with a scenario in which no norms of any kind are violated, and explicitly ask them to consider these counterfactual alternatives, CFR accounts predict we should observe these effects (Phillips et al., 2015).

Accordingly, in Experiment 2, we used a scenario that is structurally similar to the one in S&W, in which two agents use two different artifacts, bringing about some outcome. We designed an entirely norm-conforming scenario, but explicitly asked participants to consider counterfactual alternatives to either how the agent acted or what the artifact did. We predicted that asking them to consider counterfactual alternatives to the agent's action will increase causal judgments of the agent but not the artifact, and that asking them to consider alternatives to the artifact's behavior will increase causal judgments of the artifact.

1.2. Different norms, same mechanism.

The critical claim of CFR accounts is that norm violations affect causal judgments by changing the relevance of counterfactual alternatives. As noted above, CFR accounts do not expect moral norm violations to affect the relevance of alternatives relating to the behavior of inanimate artifacts. However, there are other types of prescriptive norms which should affect the relevance of counterfactual alternatives relating to artifacts: norms of *proper function*. Such norms have previously been found to affect causal judgments of artifacts (Hitchcock & Knobe, 2009; Livengood et al., 2018). However, no studies (to our knowledge) have provided direct evidence that 'malfunction' violations change the relevance of counterfactual alternatives

relating to artifacts. Furthermore, no previous work on malfunctions has looked at cases like the ones used in S&W and Experiment 2, in which different artifacts are used by different agents to produce some outcome.

Therefore, in Experiment 3, we used the same scenario that was used in Experiment 2, but instead of explicitly asking participants to consider counterfactual alternatives, we introduced either a moral norm violation or a functional norm violation. We asked participants for both causal judgments and counterfactual relevance judgments. The general prediction still holds: the two kinds of judgments should be closely correlated. However, in particular, we expect the moral violation to affect judgments of the agent but not the artifact (as S&W observed and as we replicate in Experiment 1), and the functional norm violation to affect judgments of the artifact but not the agent.

Up to this point we have incidentally confounded two features of these scenarios (as did S&W): The ontological category of a cause, and its relative proximity to the outcome. In Experiments 1-3, the artifact is always the most 'proximal' cause of the outcome, and the agent is a 'distal' cause which acts through the artifact. To ensure the patterns we observe in Experiments 1-3 were not due to some kind of specialized causal reasoning that is specific to this particular configuration of agent and artifact in a causal chain, in Experiment 4 we de-confounded these features by creating scenarios that involved exclusively artifacts or exclusively agents. We then manipulated whether it was the 'proximal' or 'distal' cause which committed a norm violation. Again, we asked for both causal and counterfactual relevance judgments, predicting the two should be closely correlated, and that the results should pattern like those in Experiments 1-3: Counterfactual relevance judgments and causal judgments of the norm-violating cause should be increased, but those of the other cause should be relatively unaffected.

2. Experiment 1

In Experiment 1, we set out to replicate the pattern of causal judgments reported by S&W and relate that pattern to an unambiguous measure of counterfactual relevance. Our stimuli were identical to S&W's stimuli. The only difference was that, prior to asking the 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. Under S&W's polysemy account, counterfactual relevance judgments should not vary based on which of these three participants were asked about, and should dissociate from causal judgments. Under the CFR account, the counterfactual relevance of alternatives involving the agent, but not the artifact, should be affected by the moral violation, and should be closely correlated with causal judgments across conditions.

2.1. Methods.

2.1.1. Participants. 610 participants (M_{age}= 37.28, SD_{age} = 12.14; 338 females, 1 unreported) from Amazon Mechanical Turk participated for modest monetary 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 a modified version of S&W's Experiment 4 with an additional DV. The overall design was 4 (norm condition) × 3 (question)

and administered fully between-subjects. The study materials were presented in Qualtrics (Qualtrics, 2005).

Participants read one of four vignettes identical to those used in S&W (see 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, so Tom decides he wants both gardeners to use only fertilizer A-X200. In all cases, however, Alex applies fertilizer A-X200 and 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; in the *Unintended norm-violation condition*, Benni believed he was 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 he uses B-Y33 instead; and in the *Deceived norm-violation condition*, Alex deliberately lies to Benni about which fertilizer he is supposed to use to get him in trouble.

Additionally, the questions that participants answered varied. As in S&W, 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 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 S&W, 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.[1]

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 S&W, 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.

2.2. Results.

As in S&W, we excluded participants who did not answer both of the check questions correctly, and analyzed the remaining 439 participants' judgments. We both qualitatively and quantitatively replicated the pattern of causal judgments observed in S&W. At the level of each condition, participants' average causal judgments in our experiment were extremely similar to that of S&W, $r = 0.950$ [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 S&W's data (Fig. 1, left panel, see also Table S1 for the complete information on the replication of the key statistical tests reported in S&W).

Next, to examine the effects of the manipulations on both causation and counterfactual relevance judgments, we categorized participants' responses as assigning causal responsibility

[1] This fixed order of questions was used because S&W 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.

(or counterfactual relevance) to (1) only the norm-violating agent, (2) both agents, or (3) only the norm-conforming agent, and 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-condition × 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-condition (LRT=13.93 [df=3], p=.003), an effect of question (LRT=73.34 [df=2], p<.001), and a norm-condition × question interaction effect (LRT=14.15 [df=6], p=.028. 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 participants' responses, judgments of the causal responsibility were highly correlated with judgments of whether it was relevant to consider alternatives to the agents' actions. 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), and also held whether participants were making judgments about agents (r=0.651 [.575, .715], p<.001), actions (r=0.262 [.157, .362], p<.001), or simply inanimate artifacts (r=0.280 [.172, .382], p<.001).

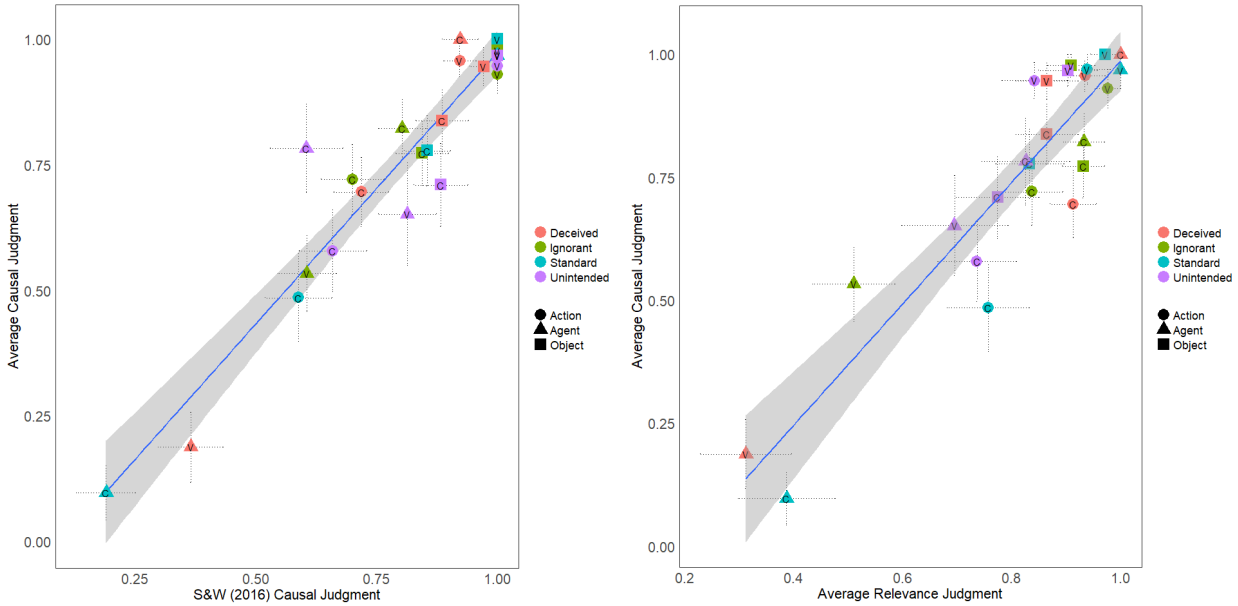

Figure 1. Depiction of the relationship between participants' causal judgments in Experiment 1 and the previous causal results from S&W (left panel) and the mean relevance judgments (right panel). Judgments related to the norm-conforming agent, action, or artifact are marked with a 'C'; Judgments related to the norm-violating agent are marked with a 'V'.

2.3. Discussion.

We successfully replicated S&W's results for causal judgments but additionally found the exact same pattern for counterfactual relevance judgments, and across all conditions, close correlations between the two. Critically, when the agent committed a moral violation, counterfactual relevance ratings for the agent increased, but counterfactual relevance ratings for the artifact did not. This demonstrates that the effects originally observed by S&W are wholly consistent with CFR accounts. Furthermore, we found a close correlation between causal and counterfactual relevance judgments both across individual judgments and averaged across conditions.

However, all of this evidence is correlational, and thus does not provide direct support for the key causal claim of CFR accounts. Moreover, it raises a further question of *why* both causal and counterfactual judgments of inanimate objects were unaffected by moral norm violations. One possibility is that causal reasoning about inanimate artifacts used as tools is simply insensitive to the consideration of counterfactual alternatives. Another possibility is that the counterfactual alternatives that become relevant in cases of *moral* norm violations typically need not involve the representation of the inanimate artifact used by the agent. We next turn to addressing this question by testing the key causal claim of CFR accounts. We do this by *directly* manipulating the relevance of counterfactual alternatives to the actions of either the agent or the artifact in a context that did not involve any norm violations.

3. Experiment 2

CFR accounts explain the pattern observed in S&W by arguing that the changes in the relevance of counterfactual alternatives for the agent (but not the artifact the agent used) *caused* participants to see the agent (but not the artifact used) as more causally responsible. This proposed mechanistic relationship between counterfactual and causal judgments is meant to be perfectly general, i.e. not specific to either agents nor moral norm violations. One should observe the same 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). In addition, we proposed that one reason counterfactuals relating to the agent did not affect causal judgments of the artifact is because these counterfactuals need not require representing or reasoning over the role played by the inanimate artifact. Therefore, if we

explicitly ask participants to consider alternatives to how the artifact functioned, we predict a corresponding increase in causal judgments of the artifact.

To test these hypotheses, we presented participants with a vignette involving a norm-conforming agent who uses an inanimate artifact leading 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.

3.1. Methods.

3.1.1. Participants. 601 participants (M_{age} = 35.96, SD_{age} = 15.58; 304 females, 2 unreported) from Amazon Mechanical Turk participated for a modest monetary compensation. Participant recruitment was again automated through TurkPrime.

3.1.2. Stimuli and procedure. This experiment used a 3 (Agent Counterfactual vs. Artifact Counterfactual vs. No Counterfactual) × 2 (Agent Question vs. Artifact Question) design. Counterfactual condition was manipulated between-subjects while Question was a within-subjects factor. The study materials were presented in Qualtrics (Qualtrics, 2005).

Participants read a vignette involving a vending machine in an academic department (see Supplementary Materials). 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, 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). 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.

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. 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 didn't produce a pencil from the vending machine. In the No Counterfactual condition, participants were simply asked to describe the story they read.

After completing this task, they rated their agreement (on a scale from 0 ('Completely disagree') to 100 ('Completely agree') with a statement that the Professor caused the problem, and separately with a statement that the red lever caused the problem. The statements were presented in counterbalanced order and on separate pages. 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.

3.2. Results.

We excluded participants who did not answer both of the check questions correctly, and analyzed the remaining 423 participants judgments. Mean ratings by question and condition are shown in Fig. 2. 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, Maechler, Bolker,

Walker, et al., 2014). This analysis revealed a main effect of Question, $\chi^2(1) = 53.135$, $p < .001$, and a main effect of Condition, $\chi^2(2) = 13.492$, $p = .001$. Critically, however, these main effects were qualified by a significant Question × Condition 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 lever functioned ($M = 24.43$, $SD = 29.12$), $t(279) = 2.27$, $p = .024$, $d = 0.272$, or when they did not 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 lever was more of a cause when they considered alternatives to the way the 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, ($M = 8.62$, $SD = 19.64$), $t(211.21) = 3.42$, $p < .001$, $d = 0.421$.

Notably, across all cases, participants' causal ratings were relatively low (below the midpoint of 50 on the agreement scale). This might indicate that participants assigned a substantial amount of causal responsibility to factors that we did not ask about, such as the administrative assistant, the student, or the other levers. However, our primary concern is the effect of counterfactuals involving a specific cause on causal judgments of that cause. To that end, only the change in ratings between conditions of the two causes we focused on are relevant.

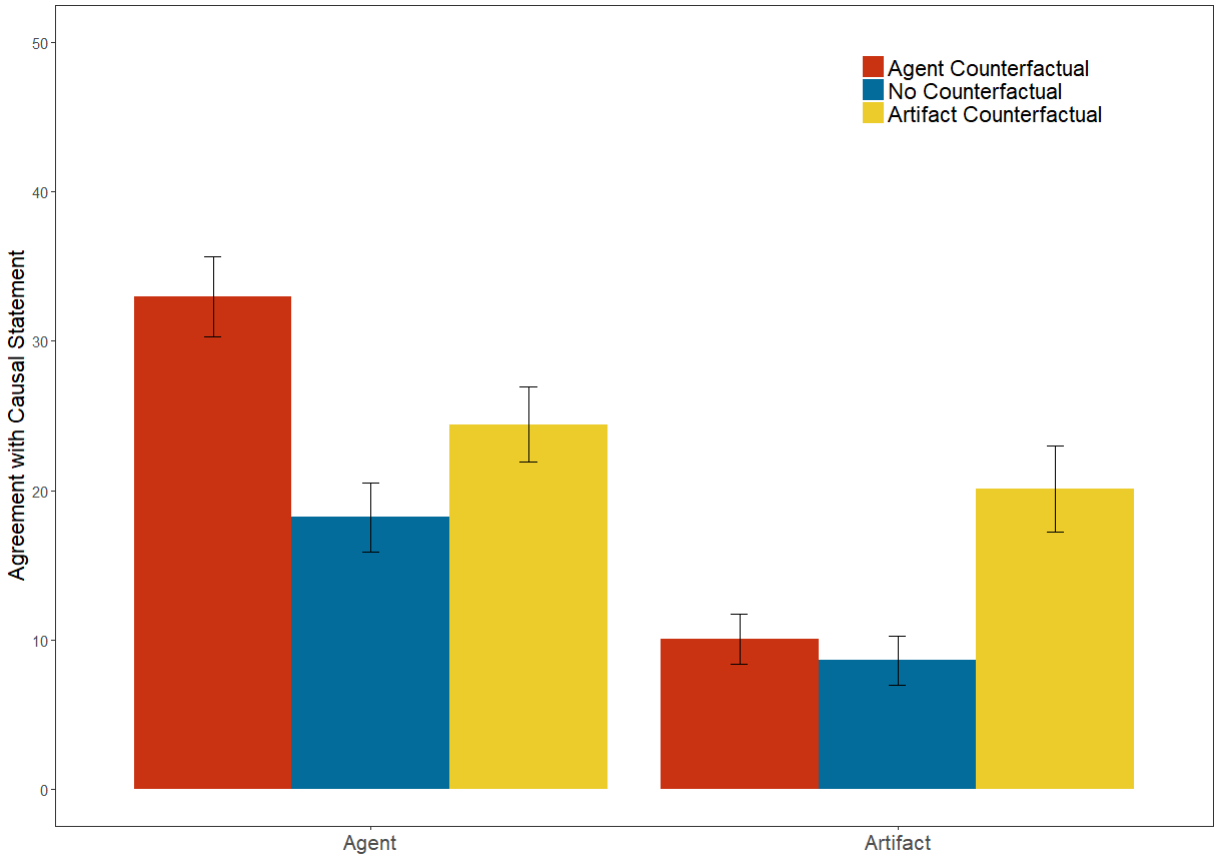

Figure 2. Average agreement ratings with the causal statement for the Agent (left) and Artifact (right) as a function of whether the counterfactual manipulation focused on the agent (red bars), there was no counterfactual manipulation (blue bars), or the counterfactual manipulation focused on the inanimate artifact (yellow bars). Error bars depict +/- 1 SEM.

3.3. Discussion.

This experiment confirmed the key mechanistic claim of CFR accounts by *directly* manipulating the relevance of counterfactual alternatives involving either the agent or artifact used, in a context that did not involve any norm violations. In particular, manipulating the relevance of counterfactual alternatives to what the artifact *did* affected causal judgments of the artifact. Moreover, manipulating the relevance of counterfactual alternatives to the agent's action

had a similar effect on causal judgments of the agent. These results jointly provide direct support for the causal mechanism suggested by the correlations observed in Experiment 1. Having validated that this underlying mechanism operates in causal reasoning about both agents and artifacts, we can now test the predictions of this mechanism in norm-violations that apply specifically to artifacts rather than agents, i.e., norms of *proper function*.

4. Experiment 3

Experiments 1 and 2 address the challenges to CFR accounts raised by S&W, and add novel evidence in support of CFR accounts in scenarios involving both agents and artifacts. We now turn our attention to demonstrating that violations of norms of proper functioning affect causal judgment by changing the relevance of counterfactual alternatives in much the same way that moral norm violations affect causal judgments of agents by changing the relevance of counterfactual alternatives.

Previous work has mostly focused on moral norm violations and statistical or descriptive norm violations (e.g., Alicke, 2001; Gerstenberg et al., 2015; Icard et al., 2017). As noted previously, moral prescriptions only apply to moral agents, and so it is unsurprising that they do not affect judgments of artifacts (Experiment 1). 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., 2018), there is currently no direct evidence that these effects are due to changes in the relevance of counterfactual alternatives. Therefore, in Experiment 3, we examined whether prescriptive norm violations that apply to inanimate artifacts affect counterfactual relevance and causal judgments of artifacts in precisely

the same way that moral norm violations affect counterfactual relevance and causal judgments of intentional agents.

4.1. Methods.

4.1.1. Participants. 403 participants (M_{age} = 34.96, SD_{age} = 11.90; 205 females, 1 unreported) from Amazon Mechanical Turk participated for a modest monetary compensation. Participant recruitment was again automated through TurkPrime.

4.1.2. Stimuli and procedure. This experiment used a 3 (Norm violation; norm-conforming vs. moral violation vs. malfunction) × 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 (see Supplementary Materials). 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 norm-conforming 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 moral violation condition was identical to the norm-conforming condition, except that the faculty are not allowed to get pencils from the machine (but

administrative assistants are allowed), and this rule was known by Prof. Smith. Lastly, the malfunction condition was identical to the norm-conforming 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 not been told 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.

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.

4.2. Results.

We excluded participants who did not answer all three of the check questions correctly, and analyzed the remaining 258 participants' judgments. 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 of these conditions were overwhelmingly met.

4.2.1. Moral check. Participants in the Moral Norm Violation condition much more strongly agreed that it was immoral for Professor Smith to pull the red lever (M = 3.69, SD = 1.71), than participants in the No Norm Violation condition (M = 5.88, SD = 1.31), $t(247.01) = -11.57$, $p < .001$, $d = 1.43$, or the Functional Norm Violation condition (M = 5.81, SD = 1.52), $t(277) = -10.95$, $p < .001$, $d = 1.31$. The No Norm Violation and Functional Norm Violation conditions did not differ significantly from one another, $t(266) = 0.41$, $p = .685$, $d = 0.04$.

4.2.2. Probability Check. Participants in all three conditions estimated the probability that the lever that gave erasers would malfunction to be well above 50%, and thus not a descriptive norm violation. Most critically, this was observed in the Functional Norm Violation condition (M = 85.32, SD = 21.39), $t(130) = 18.90$, $p < .001$. It was additionally observed in the Moral Norm Violation condition (M = 72.21, SD = 22.63), $t(126) = 11.06$, $p < .001$, and the No Norm Violation condition (M = 75.19, SD = 22.86), $t(114) = 11.82$, $p < .001$.

4.2.3. Causal judgments. Results can be found in Fig. 3, left panel. To facilitate comparison of participants' judgments, we computed a measure of participants' preference for selecting the norm-violating event as a cause. 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. 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 1. This analysis revealed a main effect of Norm condition, (LRT=71.49 [df=2], p<.001), no main effect of Causal question (LRT=0.045 [df=1], p=.832), 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 conditions. We first compared the strength of the preference for the norm-violating agent or artifact relative to the no-violation condition in each violation condition. We found, as expected, that the norm-violating artifact was more strongly preferred in the malfunction condition than the normal condition (LRT = 63.31 [df = 1], p < .001), and the norm-violating agent was more strongly preferred in the moral violation condition than the normal 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 *immoral* condition, relative to the normal 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 preference for the norm-conforming agent but not the artifact (LRT = 1.13 [df = 1], p = .288).

4.2.4. Counterfactual relevance. We next analyzed participants' judgments of the relevance of counterfactual alternatives in the same way. Results can be seen in Figure 3, 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 = .747), 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 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-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-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 salient, participants tended to prefer counterfactuals for the agent more than the artifact (LRT = 16.63 [df = 1], p < .001). When the relevant 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).

4.2.5. Relationship between causal and counterfactual judgments. Across the variations in the event asked about (Agent vs. Artifact) and Condition (Immoral vs. Malfunction vs. Normal), 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 4).

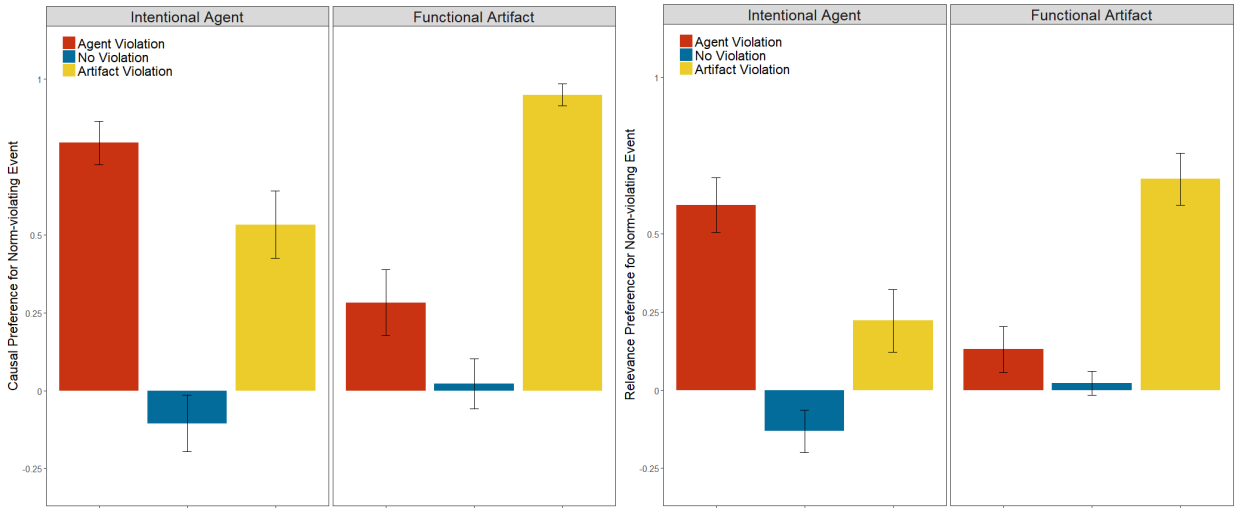

Figure 3. Average preference score for the norm-violating event in causal judgments (left plots) and counterfactual relevance judgments (right plots), as a function of whether the questions focused on intentional agents or functional artifacts (split into panels). The color of the bars indicates whether 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.

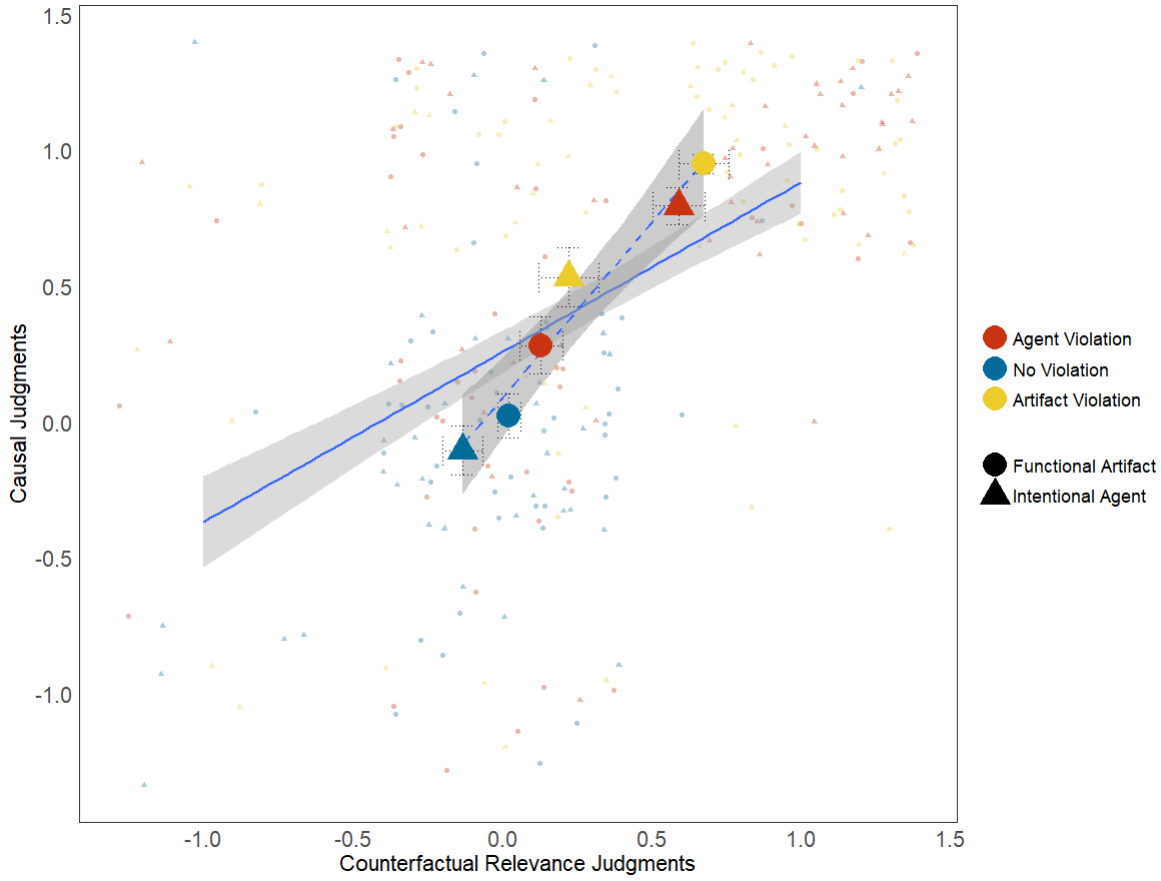

Figure 4. Relationship between causal judgments and judgments of the relevance of counterfactual alternatives. The solid line represents the linear relationship between participant-level pairs of responses (depicted by the smaller points). The dotted line represents the linear relationship between the 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.

4.3. 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 S&W by demonstrating that when the agent violates 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 violates a functional norm, judgments of the artifacts were affected more than those of the agent who used it. Most importantly, participants' causal and counterfactual judgments were clearly aligned at the level of both participants and conditions. Taken together, these patterns support CFR accounts 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 one that deserves further investigation. In a final experiment, we seek to replicate the patterns observed in this experiment in a new scenario that allows us to distinguish the effects of causal structure from those of ontological category.

5. Experiment 4

Our goal in Experiment 4 was to further investigate the unexpected effect of functional norm violations on causal judgments of intentional agents with no corresponding effect of moral violations on artifacts. A challenge in interpreting this effect is that there is a confound present in

Experiments 1-3: In all the experiments reported so far, the ontological category of different causes (whether it is an agent or an artifact) has been confounded with position in the causal structure. That is, intentional agents have always been more 'distal' from the outcome, while functional artifacts have always been 'proximal', i.e. more directly bringing about the outcome. Thus, one possibility is that the observed asymmetry arose from the agents' and artifacts' position in the causal structure, suggesting more generally that norm violations in more 'proximal' positions affect causal judgments of events in more 'distal' positions. Alternatively, it may be that the asymmetry instead arose from differences in the ontological category of the events in one location or another: perhaps proximal artifacts affect judgments of distal causes regardless of their ontological category, while distal agents do not affect judgments of proximal causes.

In Experiment 4 we de-confounded position in the causal structure and ontological category, using a new scenario in order to better understand why this asymmetry in causal judgments arose. Our approach was to isolate the role of causal structure by making scenarios with only one type of cause. To test the role of ontological category, we employed two scenarios: one in which all of the potential causes are artifacts, and one on which all the potential causes are agents. In each scenario, we used the appropriate type of norm violation for that ontological category (functional for artifacts, moral for agents), and manipulated whether the norm violation was committed by proximal cause or the distal cause (or neither violated a norm).

5.1. Methods.

5.1.1. Participants. 606 participants (M_{age} = 36.70, SD_{age} = 16.70; 290 females, 5 unreported) from Amazon Mechanical Turk participated for a modest monetary compensation. Participant recruitment was again automated through TurkPrime.

5.1.2. Stimuli and procedure. We created a novel set of vignettes for this experiment designed to have a causal structure similar to the scenarios used in Experiment 1-3, but in which the ontological category of the various causes could be varied. Each version had a "distal cause", which in some way attempted to elicit a specific action from two "proximal causes", and the actions of the proximal causes directly brought about the outcome. We manipulated two factors: the ontological category of all of the causes (all functional artifacts or all intentional agents), and which of the causes committed a category-appropriate norm violation. That is, we both manipulated the kind of norm violation (i.e., malfunction for artifacts, moral for agents) and who this norm violation was committed by (neither, distal, or proximal). The complete vignettes are included in the Supplementary Materials. The study materials were presented in Qualtrics (Qualtrics, 2005).

All vignettes involved a town that installed a new drainage system for their highway, which ran between a lake and a river. The highway had two drainage valves, a river valve and a lake valve, which are left halfway open unless instructed otherwise. In the functional artifact vignette, the town has a computer system (the distal cause) that sends a signal to two mechanisms (the proximal causes) that open or close two different valves to prevent the highway from flooding. In the intentional agent condition, a supervisor named Alex (the distal cause) sends a message to two mechanics, Billy and Sam (the proximal causes), to open or close the two valves.

In the “distal violation” condition, the distal cause is supposed to order the opening of both valves during heavy rain but ordered the lake valve to be opened and the river valve to be closed in light rain. In the functional artifact condition, this violation happened because the computer malfunctions and detects the rain as being light. In the intentional agent condition, this happens because Alex, the supervisor, is in a bad mood and wants to make people miserable by forcing the highway to close. Both proximal causes act as instructed and the highway floods.

In the “proximal violation” condition, the distal cause correctly sends a signal for both valves to open after heavy rain is detected, but the proximate cause that governs the river valve does not open it as instructed. In the functional artifact condition, this happened because the mechanism that operates the river valve malfunctions. In the intentional agent condition, Billy, the river valve mechanic, is in a bad mood and wants to make people miserable by forcing the highway to close, and so closes the river valve instead.

In the “no violation” condition, in cases of heavy rain, the distal cause was supposed to issue commands to the proximal causes to open the lake valve and close the river valve, to stop the river from overflowing. The distal cause sends the signals correctly, and both proximal causes then act as instructed. All versions concluded: “Unfortunately, this storm is unusually heavy. With one valve closed, the storm drains backed up and the highway flooded.”

Using the same scale as Experiment 2, participants asked how much they agreed with causal statements about the distal cause and the proximal cause (specifically, the proximal cause that violated a norm in the proximal violation condition, but not the other proximal cause), as well as corresponding counterfactual relevance judgments, using the same 100-point agreement scale. The order of causal and counterfactual relevance questions and whether the proximal or distal event was asked about first were both randomized.

5.2. Results.

Participants were again excluded for incorrectly answering either of two check questions, leaving 402 participants for analysis. Average causal and relevance ratings for each cause in each condition can be found in Fig. 5.

5.2.1. Causal ratings. We once again used a series of linear mixed-effects models, this time with three fixed effects, Event (within; distal vs. proximal), Condition (between; functional artifact vs. intentional agent), and Violation (between; distal event vs. proximal event vs. none). This analysis found a significant three-way interaction, $\chi^2(2) = 102.39$, $p < .001$. To give a better sense of how these results compare to the results of Experiments 1-3, we therefore conducted separate analyses of functional artifact and intentional agent conditions.

5.2.2. Functional Artifacts. There was a significant main effect of Violation, $\chi^2(2) = 54.51$, $p < .001$, a much smaller but significant main effect of Event, $\chi^2(1) = 4.36$, $p = .037$, and a significant interaction, $\chi^2(2) = 77.234$, $p < .001$. For direct comparison, we decomposed this interaction with a series of planned comparisons, as in Experiment 2.

For the proximal event, participants gave higher agreement ratings when the proximal event violated a norm than (1) when no norm was violated, $t(156) = 8.11$, $p < .001$, $d = 1.29$, or (2) when the distal event violated a norm, $t(75.24) = 5.90$, $p < .001$, $d = 1.19$. We also included new planned comparisons in light of the results of Experiment 3, which allowed us to ask whether the distal event violating a norm affected judgments of the proximal event (compared to the no-violation condition); it did not, $t(128) = 0.468$, $p = .640$, $d = .086$.

For the distal event, participants gave higher agreement ratings when the distal event violated a norm than (1) when no norm was violated, $t(119.41) = 10.76$, $p < .001$, $d = 1.804$, and

(2) when the proximal event violated a norm, $t(117.86) = 7.21$, $p < .001$, $d = 1.235$. Similar to the results found in Experiment 3, we again found that causal judgments of the distal artifact increased when the proximal artifact violated a norm (compared to when neither violated a norm), $t(155) = 2.15$, $p = .033$, $d = .344$. In short, these judgments qualitatively replicated the pattern found in Experiment 3, despite the fact that the distal event no longer involved an agent (and despite changing the DV to a continuous rating scale).

5.2.3. Intentional agents. There was a significant main effect of Violation, $\chi^2(2) = 56.28$, $p < .001$, no significant main effect of Event, $\chi^2(1) = 0.946$, $p = .331$, and a significant interaction, $\chi^2(2) = 497.56$, $p < .001$. We decomposed this interaction with the same series of planned comparisons as we conducted for the functional artifacts.

For the proximal agent, participants gave higher causal ratings when the proximal event violated a norm than (1) when no norm was violated, $t(136.88) = 27.62$, $p < .001$, $d = 4.55$, or (2) when the distal agent violated a norm, $t(132) = 22.61$, $p < .001$, $d = 3.98$. However, ratings of the proximal event were unaffected by the distal event's norm violations compared to when neither agent violated a norm, $t(111) = 0.959$, $p = .340$, $d = .181$.

For the distal event, participants gave higher agreement ratings when the distal event violated a norm than when (1) no norm was violated, $t(91.70) = 14.42$, $p < .001$, $d = 2.65$, or (2) the proximal event violated a norm, $t(130) = 24.00$, $p < .001$, $d = 4.25$. However, compared to when neither agent violated a norm, ratings were significantly *lower* when the proximal agent violated a norm, $t(97.64) = 3.53$, $p < .001$, $d = .640$. This pattern stands in stark contrast to the one observed when the proximal event involved a functional artifact (both in Experiment 3 and in the functional artifact condition in this experiment).

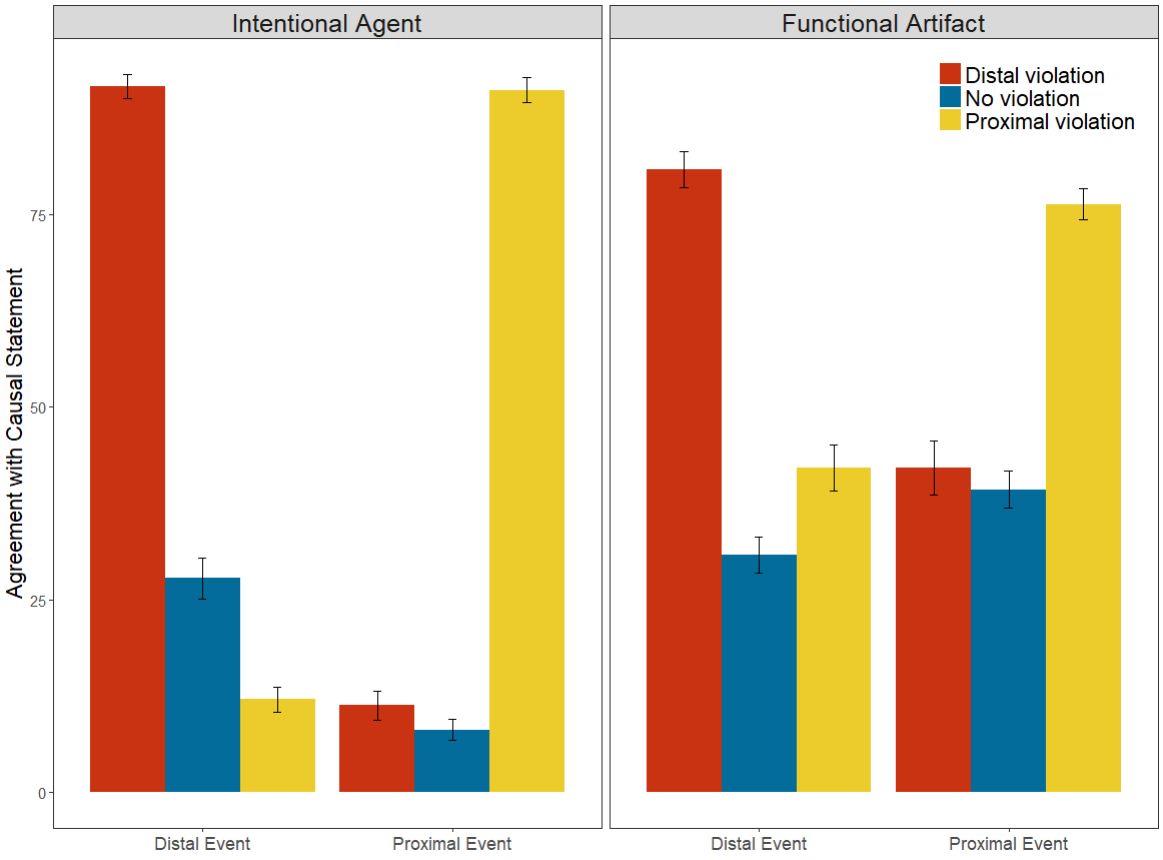

Figure 5. Average causal judgments of the distal event (left-hand bars) and the proximal event (right-hand bars) as a function of the ontological category of the event (split into panels). The color of the bars indicates whether the distal event violated a norm (red bars), no norm was violated (blue bars) or the proximal event violated a norm (yellow bars). Error bars depict +/- 1 SEM.

5.2.4. Relevance ratings. Across the variations in Event (distal vs. proximal), Condition (functional artifact vs. intentional agent), and Violation (distal event vs. proximal event vs. none), average counterfactual relevance ratings were highly correlated with average causal ratings, $r = .972$ [.899, .992], $p < .001$. Moreover, causal and counterfactual relevance judgments were also highly correlated at the level of individual judgments, $r = .760$ [.729, .788], $p < .001$

(Figure 6). Given the tight correlation, we have omitted exhaustive analyses of the judgments of counterfactual relevance (the code for these analyses is included in the online repository for those who are interested). All data, stimuli, and analysis code are available at: https://osf.io/cp2d5.

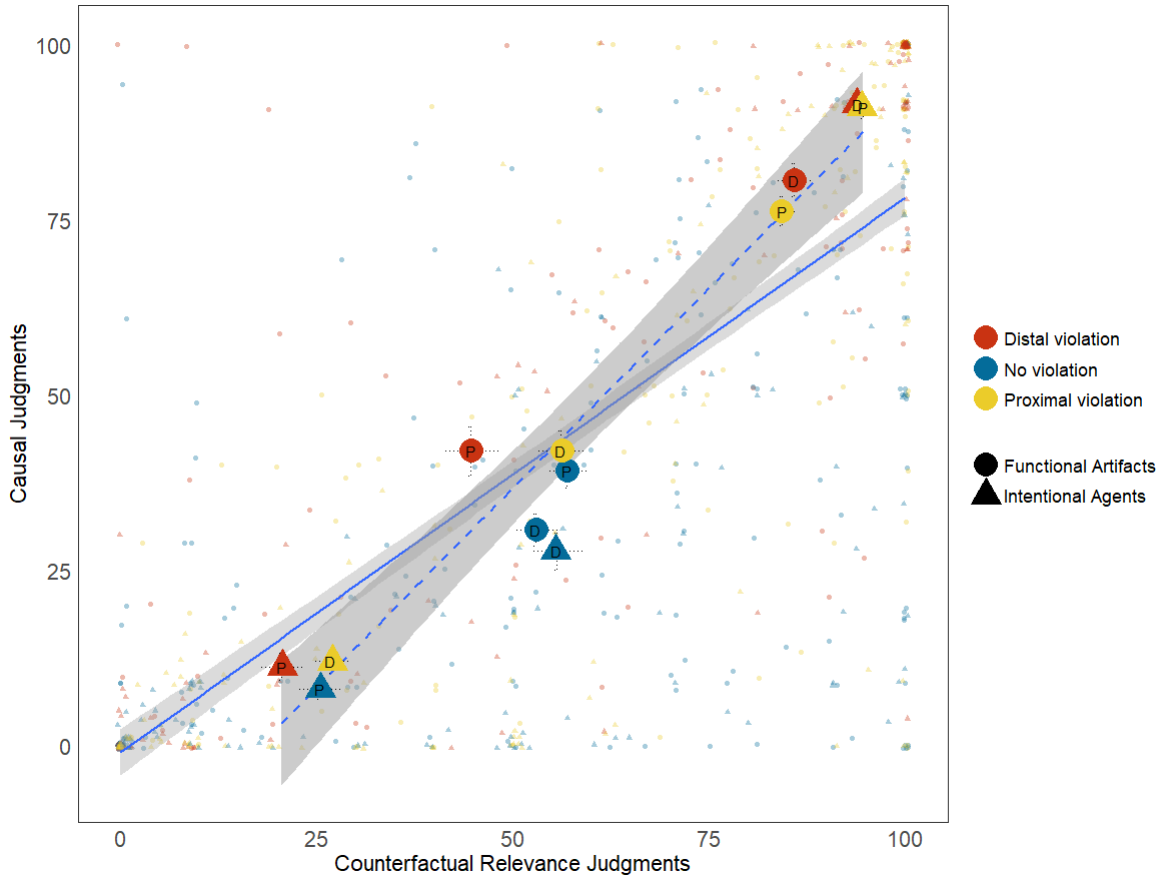

Figure 6. Relationship between causal judgments and judgments of the relevance of counterfactual alternatives. The solid line represents the linear relationship between participant-level pairs of responses (depicted by the smaller points). The dotted line represents the linear relationship between the 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 distal event violated a norm (red points), no norm was violated (blue points), or the proximal event violated a norm (yellow points). The shape of each point indicates whether the judgments

were made of functional artifacts (circles) or intentional agents (triangles). Judgments of distal events are marked by a 'D'; judgments of proximal events are marked by a 'P'. Error bars depict +/- 1 SEM.

5.3. Discussion.

In Experiment 3, we found that the malfunction of a proximal inanimate artifact increased causal judgments of a distal intentional agent. In the current experiment, we replicated that effect using only inanimate artifacts: when a proximal inanimate artifact malfunctioned, causal judgments of the distal inanimate artifact also increased. This provides clear evidence that this effect does not arise from the distal cause being an intentional agent, but instead shows that distal causes in general can be affected by proximal malfunctioning artifacts. We additionally, and unexpectedly, found that when both the proximal and distal events involved intentional agents, this effect was *inverted*, which rules out the possibility that distal intentional agents are always held responsible for the actions of proximal causes. We return to a more complete discussion of how this latter pattern ought to be understood in the general discussion.

6. General Discussion

In four experiments, we find robust support for 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 in line with them. In particular, the previously observed patterns of causal judgments are clearly predicted by the corresponding judgments of counterfactual relevance. Experiment 2 then showed that causal judgments of artifacts are affected by counterfactuals when manipulated directly. This result

provided a new demonstration that direct manipulations of counterfactual relevance (without norm violations) affect causal judgments in a reliable and predictable manner for both intentional agents and inanimate artifacts.

Experiment 3 showed that prescriptive norms 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. However, it also revealed a surprising effect wherein malfunctions of the artifacts used as tools increased causal judgments of the agent using that tool. Experiment 4 demonstrated that this unexpected effect did not arise from the ontological category of the distal agent by demonstrating that the same effect arises when the distal agent is replaced with another artifact (such as a computer), and furthermore that the effect is *inverted* when the proximal agent, the 'tool', is replaced by an intentional agent. Critically, across all of these effects, counterfactual relevance judgments were closely correlated with causal judgments, demonstrating that these patterns are not inconsistent with CFR accounts, but rather previously undiscovered phenomena that such accounts capture.

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 additional norms that have similar impact on causal and counterfactual judgments makes a 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 (Alicke et al., 2011; Samland & Waldmann, 2016). However, such an account has no obvious way of explaining (1) why moral violations have precisely the same pattern of impact on judgments of 'counterfactual relevance' which are *not* potentially polysemous, (2) 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 (3) 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.

It is also worth briefly returning to the results of the mental state manipulation in Experiment 1, which showed that mental states that mitigate moral responsibility also mitigate the effect of the norm violation on counterfactual relevance and causal judgments. S&W included this manipulation as an example of a case that moral responsibility judgments can readily explain (an ignorant or deceived norm violator is less blameworthy than a deliberate one), but (in their view) CFR accounts cannot (see also Samland et al., 2016). However, CFR accounts do not deny that agents' mental states affect their moral responsibility. In fact, CFR accounts make no claims about the process by which an event is identified as a norm violation or by which its severity is assessed, only what happens once such determinations have been made. Thus in short, we agree that the mental state of the perpetrator will likely affect the judged severity of a norm violation, or whether something is judged to be a norm violation of all, but we argue that these considerations affect *causal* judgments by changing which counterfactual alternatives are relevant.

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

6.2. What is the 'normal' counterfactual alternative?

That said, our understanding of the underlying mechanism proposed by CFR accounts is far from complete. We did find a complex and unexpected pattern of judgments that does seem to hinge on ontological category: the effect of proximal norm violations on judgments of the distal cause. The essential feature seems to be the ontological category of the proximal cause. Experiment 3 shows that, when the proximal cause is an artifact, a malfunction increases causal judgments of a distal agent. In Experiment 4, we find the same effect when the distal agent is replaced with a distal artifact. However, a proximal *moral* violation by an intentional agent *decreases* causal judgments of a distal intentional agent. (The question of how a proximal moral violation affects judgments of a distal functional artifact is left for future work.)

Some will no doubt wonder whether this new pattern we observed could be used as evidence against unified CFR accounts in which the same underlying mechanisms produce judgments about both artifacts and agents. For example, one could explain the decreased causality of the distal agent when the proximal agent violates a norm as a form of "excuse validation" (Turri & Blouw, 2015), in which people are motivated to excuse blameless actors from bad outcomes. The trouble with this approach is that standard motivated reasoning accounts

of this sort should predict the same decrease in causal judgments of the distal agent when the proximal cause is a malfunctioning inanimate artifact, since the intentional agent was not to blame. Yet, in Experiment 3 we find that the exact opposite is true. Moreover, the increase in causal judgments when the proximal cause is a malfunctioning artifact cannot be explained as a desire to assign blame to an agent over a machine, as it occurs when the distal cause is an artifact as well. But, if these effects cannot be explained by standard motivated reasoning accounts, then what does explain them?

On CFR accounts, there is a sense in which the explanation for this discrepancy is relatively straightforward: people must be considering a different set of counterfactual possibilities depending on the ontological category of the proximal cause, despite the isomorphism of the causal structure. More specifically, when the proximal cause is a malfunctioning artifact, people tend to consider alternatives that involve both the functional artifact and the distal cause that initiated the artifact's action; in contrast, when the proximal cause is an agent who acts immorally, people again tend to consider alternatives to what this agent did, but are extremely reluctant to consider alternatives that would support the causal role of the distal agent. The trouble is that we do not know exactly why these differences are occurring.

Fundamentally, one shortcoming of existing CFR accounts is that they do not have a specific account of the process by which relevant counterfactual possibilities are constructed. All versions of the CFR account hold that, in some way, people are constructing a counterfactual 'normalized' version of the event (in which the norm violations are replaced by norm-conforming actions), and 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 currently left up to intuition.

CFR accounts need not be completely insensitive to the ontological category of a cause, since, as our counterfactual relevance ratings show, counterfactual reasoning itself is not insensitive to ontological category. 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), whereas psychological events are expected to be stochastic (with many independent antecedents). One could propose a CFR account that incorporated these differences in intuitive causal structure in predicting which counterfactual possibilities are considered for malfunctions versus moral violations. For example, 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 altered (Mandel, 2003). Such an account would then explain why a malfunctioning proximal cause might affect judgments of its antecedents in Experiment 3 and the inanimate artifact condition of Experiment 4: the relevant counterfactuals were ones in which the antecedent was also changed. In contrast, in Experiment 4's agent condition, when the proximal cause was also an intentional agent who violated a moral norm, people might have considered a counterfactual which changed only the proximal agent's action. This would then lead to a causal superseding effect, as this counterfactual possibility falsifies the sufficiency of the distal agent. (Icard et al., 2017; Kominsky et al., 2015).

Of course, this proposal is speculative. Our counterfactual relevance measures are not granular enough to tell us which specific counterfactual possibilities people are considering.

Currently, we have no easy way to examine which counterfactual alternatives people consider, except by some kind of reverse inference from their causal judgments. However, such inferences risk becoming circular if they are not incorporated into a broader detailed account of how counterfactual possibilities are constructed or selected as relevant.

Exploring normality itself could provide the tools to solve this problem. Recent work on the definition of 'normality' in quantitative domains suggests that 'normal' is a combination of an ideal value and what people think the average actually is (Bear & Knobe, 2017). This may perhaps provide guidance as to what the 'normal' scenario would be for purely quantitative norm violations. Combined with other constraints on counterfactual reasoning, such as the 'nearest possible world' constraint (Lewis, 1973), it may be possible to more systematically and precisely predict which counterfactual possibilities people consider, and thereby be able to make more precise predictions about their causal judgments.

6.3. Other puzzles for CFR accounts.

A critical insight which arises in both S&W and in the current studies is that norms have a highly specific effect on causal judgments: They differentially affect causal judgments of the entities to which the norm applies and typically do not extend, or extend weakly, to other aspects of the same event. Across all of our studies, we additionally find that this pattern is replicated in participants' judgments of the relevance of counterfactual alternatives.

How to capture this sort of granularity in separating agents and artifacts in representations of counterfactuals is not something we should take for granted. CFR accounts often invoke, more or less formally, graphical modeling approaches to causal reasoning (e.g., Pearl, 2000), in which events are represented as variables that are related to each other through

structural equations. Yet, 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 which events people tend to treat as distinct causal variables, or how such variables are selected to populate the model in the first place (*cf*. Halpern & Hitchcock, 2010; Goodman, Mansinghka, & Tenenbaum, 2007). Our data provide some limited insight: The agent's decision to commit the norm violation, the use of the artifact, and the behavior of the artifact all seem to be separable in the causal model participants are reasoning over. However, we could just as easily have predicted that all three would be represented by a single causal variable (e.g., Waldmann & Mayrhofer, 2016), and have been no less justified. This issue of how we carve events into causal variables is an area severely in need of exploration in future work.

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 the Harvard's Moral Psychology Research lab for feedback on this work. JFK was supported by NIH grant F32HD089595.

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.

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.

Bear, A., & Knobe, J. (2017). Normality: Part descriptive, part prescriptive. *Cognition*, *167*, 25-37. doi:10.1016/j.cognition.2016.10.024

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 Conference 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

Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. *The British Journal for the Philosophy of Science*, *56*(4), 843–887. http://doi.org/10.1093/bjps/axi147

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

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

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. *The Journal of Philosophy*, *70*(17), 556-567. doi:10.2307/2025310

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.

Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303–332. http://doi.org/10.1016/j.cogpsych.2010.05.002

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

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. http://doi.org/10.1016/j.cognition.2015.08.001

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

Phillips, J., & Shaw, A. (2015). Manipulating morality: Third-party intentions alter moral judgments by changing causal reasoning. *Cognitive Science*, 39(6), 1320-1347.

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. http://doi.org/10.1037/xge0000138

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

Sinclair, R. C., Hoffman, C., Mark, M. M., Martin, L. L., & Pickering, T. L. (1994). Construct Accessibility and the Misattribution of Arousal: Schachter and Singer Revisited. *Psychological Science*, *5*(1), 15-19. doi:10.1111/j.1467-9280.1994.tb00607.x

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

Turri, J., & Blouw, P. (2015). Excuse validation: a study in rule-breaking. *Philosophical Studies*, *172*(3), 615-634. doi:10.1007/s11098-014-0322-z

Waldmann, M. R., & Mayrhofer, R. (2016). Chapter Three - Hybrid Causal Representations. *Psychology of Learning and Motivation*, *65*, 85-127.