Cognition 137 (2015) 196–209

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Causal superseding

Jonathan F. Kominsky [a,*], Jonathan Phillips [a], Tobias Gerstenberg [b], David Lagnado [c], Joshua Knobe [a]

[a] Yale University, United States
[b] Massachusetts Institute of Technology, United States
[c] University College London, United Kingdom

ARTICLE INFO

Article history:
Received 23 July 2014
Revised 21 January 2015
Accepted 27 January 2015

Keywords:
Causal reasoning
Counterfactuals
Morality
Superseding

ABSTRACT

When agents violate norms, they are typically judged to be more of a cause of resulting outcomes. In this paper, we suggest that norm violations also affect the causality attributed to other agents, a phenomenon we refer to as "causal superseding." We propose and test a counterfactual reasoning model of this phenomenon in four experiments. Experiments 1 and 2 provide an initial demonstration of the causal superseding effect and distinguish it from previously studied effects. Experiment 3 shows that this causal superseding effect is dependent on a particular event structure, following a prediction of our counterfactual model. Experiment 4 demonstrates that causal superseding can occur with violations of non-moral norms. We propose a model of the superseding effect based on the idea of counterfactual sufficiency.

1. Introduction

In the 1870 case of *Carter v. Towne*, the court faced an intriguing causal question. The defendant sold gunpowder to a child. The child's mother and aunt hid the gunpowder, but in a location that they knew the child could find and access. The child found the gunpowder and was injured. The court judged that the defendant could not be considered to be the cause of the child's injuries, because of the negligence of the mother and aunt (Hart & Honoré, 1985, pp. 281–282).

This case leaves us with an interesting puzzle about causal reasoning. The question before the court was not whether the mother and aunt caused the outcome; it was whether the defendant caused the outcome. Yet the court determined that the fact that the actions of the mother and aunt were negligent had some effect on the causal relationship between the defendant's actions and the outcome. This suggests a broader phenomenon of causal reasoning: the extent to which one agent is perceived to have caused an outcome may be affected not only by his or her own actions, but also by the normative status of other people's actions. We refer to this as 'causal superseding'.

It is well-established that judgments of norm violations, such as moral norm violations, can affect causal judgments. An agent who acts in a way that is judged to be morally wrong is seen as more causal than an agent whose actions conform with moral norms (e.g., Alicke, 1992). Recent work has suggested that, rather than being about morality specifically, these effects are rooted in the normality of an agent's actions, i.e., how much they diverge from prescriptive or statistical norms (Halpern & Hitchcock, 2014; Hitchcock & Knobe, 2009; but see Alicke, Rose, & Bloom, 2011). However, most of the work to date has focused on how the normality of an agent's actions affects that agent's own causality, not anyone else's. The present experiments aim to demonstrate and

[*] Corresponding author at: Department of Psychology, Yale University, Box 208205, New Haven, CT 06520-8205, United States. Tel./fax: +1 (203) 432 6451.
E-mail address: jonathan.kominsky@yale.edu (J.F. Kominsky).

http://dx.doi.org/10.1016/j.cognition.2015.01.013
0010-0277/© 2015 Elsevier B.V. All rights reserved.

explore the causal superseding effect suggested by the intriguing case of *Carter v. Towne*.

1.1. Describing causal superseding

Before discussing how the phenomenon of causal superseding may provide helpful insight into causal reasoning more generally, it is worth considering how causal superseding is related to previous research. In general, there has been relatively little research suggesting that causal judgments about one agent are affected by aspects of some other independent agent. That the actions of one person can have an influence on causal judgments about another person has been demonstrated in the relatively under-discussed research on causal chains where multiple agents collectively contribute to the occurrence of some harm (Fincham & Roberts, 1985; Fincham & Shultz, 1981; Gerstenberg & Lagnado, 2012; Lagnado & Channon, 2008; McClure, Hilton, & Sutton, 2007; Spellman, 1997; Wells & Gavanski, 1989). Among other findings, these studies report a pattern whereby the first agent in the causal chain was judged to be less of a cause of the harm that eventually occurred when the second (more proximal) agent acted *voluntarily*, rather than *involuntarily*. The explanation offered for this effect was that the voluntariness of the proximal agent's action 'broke' the perceived causal chain between the first agent and the outcome. This effect differs from the superseding effect suggested by *Carter v. Towne*. In that case, it was not the voluntariness of the aunt and mother's actions, but the *negligence* of their actions that prevented the defendant from being a cause of the child's injuries. Another closely related line of work investigated the role of 'mutability' (the ease with which the cause can be imagined to have been different) and 'propensity' (the likelihood that the effect would occur if the cause was present) in causal judgments (McGill & Tenbrunsel, 2000). This study found that one causal factor is seen as weaker when another causal factor is more mutable, though only when the mutable cause is also very likely to bring about the outcome.

Here, we specifically focus on the role of norm violations and consider their impact on causal judgments across a number of different causal structures. However, even focusing on norm violations, we also wish to acknowledge two alternative explanations for the phenomenon we investigate, one informed by intuition and the other based on existing and well-supported motivational theories.

First, one might intuitively think that "there is only so much causality to go around," and it is already known that when an agent does something that is morally wrong or otherwise in violation of some norm, that agent's causality is increased (Alicke, 1992; Hitchcock & Knobe, 2009). Then, if the norm-violation of one agent's action increases that agent's causality, it follows under this intuition that some other factor's causality will have to be reduced. Though this explanation might seem compelling at first, there is already empirical evidence that causal responsibility is not generally a zero-sum judgment (Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2014; Lagnado, Gerstenberg, & Zultan, 2013; Teigen & Brun, 2011). For example, when an outcome was brought about by a collection of causes that were each individually necessary for its coming about, then each cause was judged as fully responsible (Lagnado et al., 2013; Zultan, Gerstenberg, & Lagnado, 2012). Thus, while perhaps intuitively attractive, we do not believe this explanation can account for causal superseding.

Second, it is already known that people's causal judgments can be impacted by motivational factors. For example, a series of studies have found that people's judgments are often distorted by "blame validation" (Alicke, 1992, 2000; Alicke, Buckingham, Zell, & Davis, 2008; Lagnado & Channon, 2008): A motivational bias to assign causality to people who are blameworthy, with only minimal regard for their actual causal status. Subsequent work has extended this account to include "excuse validation" (Turri & Blouw, 2014): The motivation *not* to assign causality to individuals whom we do *not* feel are blameworthy. For example, if a driver is speeding because of an accelerator malfunction and gets into a lethal accident, we might be disinclined to regard the driver as a cause of the accident because her actions are blameless. This basic idea could then be used to explain causal superseding. If one agent does something morally wrong and is therefore seen as the one who is to blame for the outcome, people could be motivated to exculpate all other agents from blame, and may accordingly reduce the extent to which they are seen as causing the outcome.

The latter explanations draw on claims that have already received extensive support in the existing empirical literature, and we do not mean to call these empirical claims into question here. Instead, we simply provide experimental evidence for causal superseding that requires an importantly different kind of explanation. Thus, the present research goes beyond what has been demonstrated in previous work, but is not incompatible with it.

1.2. A counterfactual account of causal superseding

We propose an account of the superseding effect based on counterfactual reasoning. According to this account, the effects of valence on causal judgments are mediated by counterfactual reasoning. This account follows two key claims: First, counterfactual reasoning affects causal judgment; second, moral valence affects counterfactual reasoning. We will explore each of these claims in turn.

1.2.1. Counterfactual reasoning and causal judgment

There are many accounts of how counterfactual reasoning interacts with causal judgment (e.g., Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2014; Lewis, 1973; Petrocelli, Percy, Sherman, & Tormala, 2011; but see Mandel, 2003). We focus here on an aspect of the relationship between counterfactuals and causation that has been referred to as sensitivity (or robustness) of causation (Hitchcock, 2012; Knobe & Szabó, 2013; Lombrozo, 2010; Woodward, 2006).

Existing work on counterfactual theories of causation suggests that people regard an event as a cause of the outcome when it satisfies two counterfactual conditions, 'necessity' and 'sufficiency' (e.g., Pearl, 1999; Woodward, 2006). Take the causal relationship "A caused B". Roughly speaking, this relationship would have the following necessity and sufficiency conditions:

> *Necessity:* If A had not occurred, B would not have occurred.
> *Sufficiency:* If A had occurred, then B would have occurred.

Our focus here will be on the second of these conditions – sufficiency – and on the role it plays in ordinary causal cognition.

Woodward (2006) defines a property he calls 'sensitivity' to describe the robustness of a causal relationship. A causal condition (necessity or sufficiency) is 'sensitive' if it would cease to hold if the background conditions were slightly different. By contrast, a causal condition is 'insensitive' if it would continue to hold even if the background conditions were substantially different. Woodward argues that when the sufficiency condition is highly sensitive, people will be reluctant to attribute causation.

To give a concrete example, consider two sufficiency conditions: "If a lit match had been put near gunpowder, the gunpowder would have exploded" and "If you had manufactured fireworks, the child would have been injured." The first statement is extremely insensitive, or robust. There are a large number of things that you can change about the state of the world, but the sufficiency statement will still hold true. That is not to say that there are no changes to the background conditions that would render the statement false, but they are relatively non-obvious or non-salient. In contrast, the second sufficiency statement is more sensitive, because there are a large number of immediately salient counterfactual possibilities that would render it false. For example, the child may not be able to purchase the fireworks, or use them with supervision, etc.

This claim about the importance of sufficiency is the first piece of our account of causal superseding. In the case of *Carter v. Towne,* for example, the defendant's action was only sufficient to bring about the outcome because the mother and aunt happened to act negligently. If the mother and aunt had not acted negligently, then even if the defendant had performed exactly the same action, the outcome would not have come about. It is for this reason, we claim, that people are somewhat disinclined to regard the defendant's action as having caused the child's injuries. Certain facts about the child's guardians make the relationship between the defendant and the outcome *sensitive*.

1.2.2. Moral valence, norm violations, and counterfactuals

We now need to add a second piece to the puzzle. We noted above that a relationship could be considered 'sensitive' to the extent that it would not have held if the background circumstances had been slightly different. Yet, there will always be *some* way that the background circumstances could have been different such that sufficiency would no longer hold. For example, suppose that someone said, "The gunpowder only ignited because it was not covered in water. If it were covered in water, the match would not have been sufficient." Though this counterfactual claim is surely correct, there seems to be some important sense in which it is *irrelevant* – not even worth thinking about. If we want to understand the notion of sensitivity, we need to say more about this issue, providing a sense of how to determine whether a given counterfactual is relevant or not.

Fortunately, there is a substantial body of research on counterfactual reasoning (for reviews, see Byrne, 2005; Kahneman & Miller, 1986). This research has used a variety of techniques to explore the factors that make people regard counterfactuals as more or less relevant, and we can turn to this literature for insights into the present question.

Although research on counterfactual reasoning has uncovered a variety of notable effects, we focus here on two principal findings. First, studies show that *likelihood* judgments play a role in people's intuitions about which counterfactuals are relevant and which are not (Byrne, 2005; Kahneman & Tversky, 1982). When something unlikely occurs in the actual world, people tend to regard as relevant the counterfactuals that involve something more likely occurring. Second, studies show that *moral* judgments can influence people's intuitions about the relevance of counterfactuals (McCloy & Byrne, 2000; N'gbala and Branscombe, 1995). When an agent performs a morally bad action, people tend to regard as relevant the counterfactual in which this agent did not perform the morally bad action.

To unify these two findings, we can say that people's intuitions about the relevance of counterfactuals are affected by violations of *norms* (Hitchcock & Knobe, 2009). In some cases, an event is seen as unlikely (and hence violates a statistical norm); in other cases, an event is seen as morally wrong (and hence violates a prescriptive norm). Even though these two types of norm violation are in many ways quite different, they appear to have precisely the same effect on counterfactual reasoning. Thus we can formulate a more general principle, which should apply across both types of norm violation. The general principle is: when an event in the actual world is perceived as violating a norm, people tend to regard as relevant the counterfactuals in which the norm-violating event is replaced by a norm-conforming event.

This claim about the impact of norm violations on counterfactual thinking has played a role in some existing theoretical work in causal cognition (Halpern & Hitchcock, 2014; Hitchcock & Knobe, 2009; Knobe & Szabó, 2013), and it forms the second piece of our explanation of causal superseding.

1.2.3. The counterfactual account of causal superseding

Putting these ideas together, we end up with a counterfactual account of causal superseding, which we refer to as the *counterfactual sufficiency account*. Take the causal claim, "The defendant selling gunpowder to the child caused the child's injuries." The sufficiency condition for this claim reads as follows: "If the defendant had sold gunpowder to the child, then the child would have been injured." Now suppose that sufficiency holds only because the mother and aunt negligently hid the gunpowder where the child could find it. Since this act violates a norm, people will tend to regard as highly relevant the possibility in which the gunpowder is put somewhere that the child could not find it. In that possibility, the defendant's action is not sufficient, so the negligent actions of the mother and aunt make the defendant's sufficiency more sensitive. Thus, the defendant is regarded as less of a cause of the outcome, or in other words, is superseded.

By contrast, suppose that the mother and aunt's actions did not violate a norm, but nonetheless the defendant was sufficient only because of their (normative) actions. Then it might still be true that sufficiency would not have held if the mother and aunt had acted differently, but the possibility in which they acted differently would be regarded as less relevant and the sufficiency of the defendant's action would not be seen as especially sensitive to background circumstances. Instead, it might be felt that the defendant's action would have been sufficient for bringing about the negative outcome in all of the possibilities that are genuinely worth considering.

Putting this point more abstractly: Suppose that there are two agents, A and B, such that the outcome would not have arisen if either of these agents had acted differently. When agent A violates a norm, it makes possibilities in which they do *not* violate that norm very relevant. If the sufficiency condition for agent B is not met in those possibilities, the sufficiency of the causal link between agent B and the outcome becomes sensitive. Because the sufficiency of that causal link is sensitive, agent B is seen as less of a cause of the outcome. This model is represented in Fig. 1.

1.3. Predictions of the counterfactual sufficiency model

Our account of causal superseding makes several specific, novel, and testable predictions. The first novel prediction is that causal superseding should occur even for outcomes that are in no way bad. This goes beyond, but does not contradict, motivational accounts (e.g., Turri & Blouw, 2014). If you are highly motivated to justify the conclusion that an agent is not blameworthy, you can do so by making a causal judgment of the form, "This agent did not cause the bad outcome." However, that same logic does not apply when the outcome is not bad. In such a case, you might still be motivated to justify the conclusion that the agent is not blameworthy, but you could not justify that conclusion by making a causal judgment of the form, 'This agent did not cause the neutral (or good) outcome.' Such a judgment would not directly help to show that the agent was not blameworthy. Existing work on motivational biases in causal cognition has used precisely this logic to show that certain effects are indeed the product of motivation (Alicke et al., 2011). In contrast, the counterfactual account does not require that the outcome is bad in order for superseding to occur. From the standpoint of the counterfactual account, the relevant component is the norm violation of the superseding actor (A in Fig. 1), not the valence of the outcome.

The second prediction is not about when superseding should occur, but rather when it should not. The counterfactual account does not treat the assignment of causality to different actors as a zero-sum problem. Our account predicts that superseding should occur only when the sufficiency of the superseded actor is threatened. If that is not the case, the wrongness of one agent's action should not decrease the other's judged causality. Consider a situation in which an outcome happens if either A *or* B (or both) act. Here, no matter whether or not A acts, B's action is sufficient for bringing about the outcome. In this situation the sensitivity of B's sufficiency for the outcome is independent of A, and therefore we predict that varying the normality of A's action will not affect causal judgments of B's action.

Finally, the third prediction is that superseding should arise for any norm violation, not just for violations of moral norms. The key role that moral valence plays in the counterfactual account is that of making certain counterfactual possibilities more relevant, and those possibilities make the sufficiency of the superseded actor sensitive. Previous work has suggested that violations of other norms, such as purely statistical norms, should make counterfactual

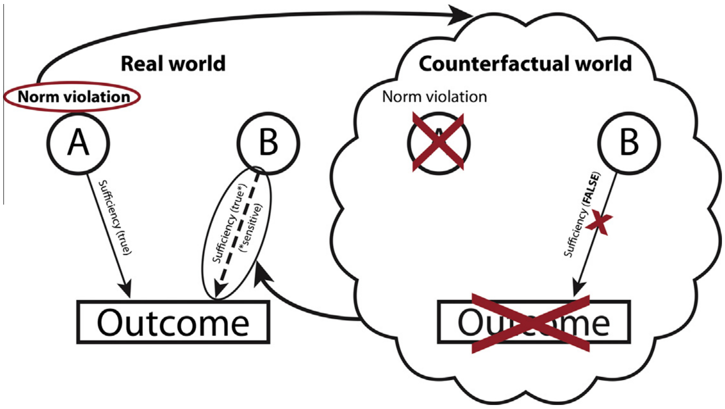

Fig. 1. The counterfactual model of causal superseding. A's norm violation (Real world) leads people to consider the counterfactual possibility in which that norm violation did not occur (Counterfactual world). The relationship between B's action and the outcome is sensitive to the extent that the outcome would not have occurred in the counterfactual world in which A's norm violation wouldn't have taken place. The more sensitive the sufficiency relationship between B's action and the outcome, the less causally responsible is B's action for the outcome: A's action supersedes B's causality.

possibilities more relevant in the same way (Kahneman & Tversky, 1982), so violations of these norms should yield similar superseding effects.

We test these three predictions in four experiments. Experiments 1 and 2 investigate the role of outcome valence in causal superseding. Experiment 3 tests the second prediction, concerning cases in which superseding should not occur because each actor is independently sufficient. Finally, Experiment 4 investigates whether superseding arises not only for violations of moral norms but also for violations of statistical norms.

2. Experiment 1

In the first experiment, we aimed to demonstrate the basic phenomenon of causal superseding. We constructed a scenario with two agents whose actions combine in a conjunctive way to bring about a neutral outcome. One agent, whom we will call the 'fixed' agent, always acted in the same way. Her actions were always morally neutral. The second agent, whom we will call the 'varied' agent, did something either morally neutral or morally wrong, depending on condition. In order to validate our manipulation and verify the neutrality of the outcome, we asked participants to rate how good or bad each agent's actions were after making causal ratings, as well as how good or bad the outcome was. The counterfactual sufficiency account predicts that the fixed agent should be seen as less causal when the varied agent's actions are morally wrong, and that this effect should arise regardless of the valence of the outcome.

2.1. Methods

2.1.1. Participants

60 participants were recruited via Amazon Mechanical Turk and paid $0.20 each for completing the survey.

2.1.2. Materials and procedure

Two vignettes were created featuring a varied agent (Bill) and a fixed agent (Sue). The fixed agent's actions remained constant in both vignettes. The moral wrongness of the varied agent's actions were manipulated between conditions (see Table 1).

In all conditions, participants were asked to rate on a 1 (strongly disagree) to 7 (strongly agree) scale how much they agreed with each of the following two sentences: "Sue caused them to possess the paired set of bookends" (the fixed agent) and "Bill caused them to possess the paired set of bookends" (the varied agent). Questions were presented in random order.

After the causal ratings, participants were asked to rate the valence of each agent's actions, as well as the outcome, on a separate page on which they could not see their previous ratings or the vignette. Participants were asked: "How good or bad is it that Sue bought the left-side Bartlett bookend from the antique store", "How good or bad is it that Bill [bought/stole] the right-side Bartlett bookend from his friend" (depending on condition), and "How good or bad is it that Bill and Sue have a paired set of Bartlett bookends". Participants made their ratings on a 1–7 scale, with "very bad" (1) and "very good" (7) at the endpoints and "neither good nor bad" (4) at the midpoint. The three questions were presented in randomized order on the same page.

2.2. Results and discussion

We evaluated the effect of the moral valence manipulation on causal ratings for each agent independently, as well as valence ratings for each agent and the outcome.

2.2.1. Causal ratings

The agreement ratings for causal questions can be found in Fig. 2. Replicating many previous studies, agreement ratings for the varied agent (Bill) were higher when he violated a norm ($M = 5.97$, $SD = 1.564$) than when he

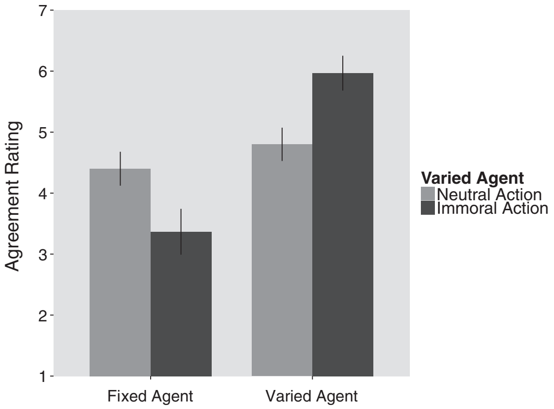

Fig. 2. Mean agreement ratings with the causal statements about the fixed agent and the varied agent as a function of the morality of the varied agent's action in Experiment 1. Error bars depict SE mean.

Table 1
Vignettes for Experiment 1.

(1) Background: Bill's wife, Sue, is out of town for the weekend. She leaves Bill a message that says, 'I just saw this marvelous bookend. It's called a Bartlett bookend. So pretty! I'm going to go back tomorrow and get one. It will be perfect for the left side of our bookshelf'	
(2a) Morally good: Bill goes and visits his friend. Bill and his friend talk for a while, and when Bill asks if his friend is willing to sell the bookend, his friend is happy to sell it. Bill makes an offer, but his friend insists on him not paying so much. Finally, Bill buys the right-side Bartlett bookend from his friend and goes home	*(2b) Morally bad:* Bill goes and visits his friend. Bill and his friend talk for a while, and when Bill asks if his friend is willing to sell the bookend, his friend tells him it's a precious heirloom and he can't part with it. Bill waits until later when his friend is in the bathroom, and slips the bookend into his bag. Finally, Bill leaves his friend's house with the stolen right-side Bartlett bookend in his bag
(3) Outcome: Then the next day, Sue goes and buys the left-side Bartlett bookend. So, when Sue got home, they had the paired set of bookends	

did not ($M = 4.80$, $SD = 1.495$), $t(58) = 2.953$, $p = .005$, $d = .764$.

For the fixed agent (Sue), we found a clear causal superseding effect. Agreement ratings for the fixed agent were lower when the varied agent violated a norm ($M = 3.37$, $SD = 2.059$) than when he did not ($M = 4.40$, $SD = 1.522$), $t(58) = -2.210$, $p = .031$, $d = .568$.

2.2.2. Valence ratings

One participant failed to give a valence rating for the fixed agent's actions, but their data were included in all other analyses. As expected, the varied agent's actions were seen as significantly worse when he stole the bookend ($M = 1.30$, $SD = .702$) than when he bought it ($M = 5.53$, $SD = 1.137$), $t(58) = -17.355$, $p < .001$, $d = 4.477$. This validates our manipulation as a violation of a moral norm. As expected, the fixed agent's actions were rated equally good whether the varied agent violated a norm ($M = 5.10$, $SD = 1.205$) or not ($M = 5.23$, $SD = 1.278$), $t(57) = -.401$, $p = .69$.

Unexpectedly, the outcome was seen as significantly worse when the varied agent violated a norm ($M = 2.93$, $SD = 1.388$) than not ($M = 5.70$, $SD = 1.022$), $t(58) = -8.791$, $p < .001$, $d = 2.272$. This suggests that participants were re-interpreting the outcome, which did not change between conditions, based on the moral valence of the varied agent's actions.

To sum up, we demonstrated the predicted causal superseding effect. When the varied agent performed a morally bad action, the fixed agent was seen as less causal. However, because the outcome was seen as bad in the conditions where the varied agent performed a morally bad action, we did not succeed in testing the hypothesis that the causal superseding effect would arise even in cases where the outcome was not itself seen as bad.

3. Experiment 2

In Experiment 2, we sought to address an alternative explanation for the causal superseding effect observed in Experiment 1. While we intended the outcome in Experiment 1 to be neutral in all conditions, participants did not see it that way. Accordingly, a better demonstration would be to explicitly manipulate the outcome and show that the superseding effect does not interact with the outcome manipulation. In Experiment 2, we thus manipulated both the valence of the outcome and the valence of one of the agent's actions.

3.1. Methods

3.1.1. Participants

120 participants were recruited via Amazon Mechanical Turk and paid $0.20 each for completing the survey.

3.1.2. Materials and procedure

We created four vignettes, which were lightly modified versions of the vignettes used by Reuter, Kirfel, van Riel, and Barlassina (2014). The vignettes involved two people, Billy and Suzy, who work at the same company and have their offices in separate rooms. The key event involved Billy and Suzy simultaneously accessing a central computer at 9 am. We manipulated two elements of this story. First, Billy (the varied agent) was either prohibited from accessing the computer at 9 am, or allowed to. Second, two people simultaneously accessing the computer either had an unexpected positive effect (deleting e-mails containing dangerous viruses) or unexpected negative effect (deleting e-mails containing important customer information). This led to a 2 (varied agent valence) × 2 (outcome valence) design, which was administered between-participants. The vignettes can be found in Table 2.

At the end of the vignette, participants were asked analogous questions to those used in Experiment 1. In all conditions, participants were asked to rate on a 1 (strongly disagree) to 7 (strongly agree) scale how much they agreed with each of the following two sentences: "Suzy caused [outcome]" (the fixed agent) and "Billy caused [outcome]" (the varied agent), with the outcome adjusted depending on the outcome valence condition. These questions were presented in random order.

After answering these causal agreement questions, participants were asked to rate the valence of each agent's action and the valence of the outcome on a separate page, using the same scales as Experiment 1. These three questions were again presented in random order.

3.2. Results

3.2.1. Causal ratings

The agreement ratings for the causal questions can be found in Fig. 3. We conducted two separate 2 (varied agent valence) × 2 (outcome valence) ANOVAs for agreement ratings of the varied and fixed agent. For the varied agent, there was a strong effect of varied agent valence, with higher agreement ratings when the varied agent violated a norm ($M = 5.98$, $SD = 1.477$) than not ($M = 3.97$, $SD = 2.077$), $F(1,116) = 37.618$, $p < .001$, $\eta_p^2 = .096$. There was no effect of outcome valence, $F(1,116) = .270$, $p = .6$, and no interaction, $F(1,116) = 1.826$, $p = .179$.

For the fixed agent, we once again found the causal superseding effect. Agreement ratings for the fixed agent were lower when the varied agent violated a norm ($M = 2.13$, $SD = 1.851$) than not ($M = 4.12$, $SD = 2.044$), $F(1,116) = 34.064$, $p < .001$, $\eta_p^2 = .227$. There was also a main effect of outcome valence, with slightly lower ratings when the outcome was negative ($M = 2.66$, $SD = 2.157$) than positive ($M = 3.55$, $SD = 2.129$), $F(1,116) = 7.478$, $p = .007$, $\eta_p^2 = .061$.

Crucially, the interaction between outcome valence and varied agent valence was not significant, though it was marginal, $F(1,116) = 3.158$, $p = .078$. To conclusively determine whether outcome valence impacted the causal superseding effect, we turned to participant judgments of the valence of the outcome.

3.2.2. Valence ratings

While the critical valence rating is the outcome valence, we also analyzed the valence ratings for each agent's actions. For the varied agent, there were strong main effects of the varied agent's action valence ($p < .001$,

Table 2
Vignettes for Experiment 2 (closely based on vignettes used by Reuter et al., 2014).

(1a) Outcome good: Billy and Suzy work for the same company. They work in different rooms and both of them sometimes need to access the central computer of the company. Nobody at the company is aware that if two people are logged into the central computer at the same time, some spam e-mails containing dangerous viruses are immediately deleted from the central computer	*(1b) Outcome bad:* Billy and Suzy work for the same company. They work in different rooms and both of them sometimes need to access the central computer of the company. Nobody at the company is aware that if two people are logged into the central computer at the same time, some spam e-mails containing important customer information are immediately deleted from the central computer
(2a) Morally neutral: In order to make sure that two people are available to answer phone calls during designated calling hours, the company issued the following official policy: Billy and Suzy are both permitted to log into the central computer in the mornings, and neither of them are permitted to log into the central computer in the afternoons	*(2b) Morally wrong:* In order to make sure that one person is always available to answer incoming phone calls, the company issued the following official policy: Suzy is the only one permitted to log into the central computer in the mornings, whereas Billy is the only one permitted to log into the central computer in the afternoons. Billy is not permitted to log into the central computer in the morning
1a (con't) Outcome good: Today at 9am, Billy and Suzy both log into the central computer at the same time. Immediately, some work e-mails containing dangerous viruses are deleted from the central computer	*1b (con't) Outcome bad:* Today at 9am, Billy and Suzy both log into the central computer at the same time. Immediately, some work e-mails containing important customer information are deleted from the central computer

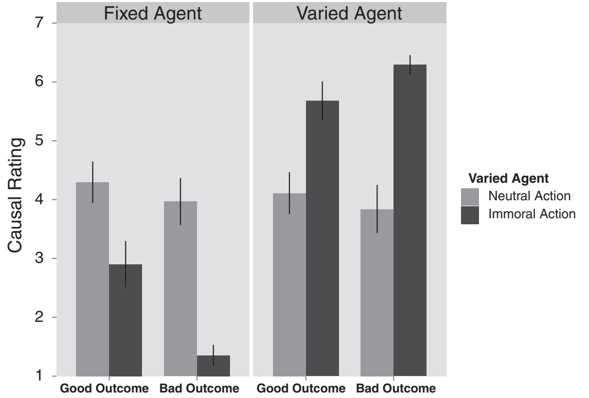

Fig. 3. Mean agreement ratings with the causal statements about the fixed agent (left panel) and the varied agent (right panel) as a function of outcome valence and the morality of the varied agent's action in Experiment 2. Error bars depict SE mean.

$\eta^2_p = .186$) and outcome valence ($p < .001$, $\eta^2_p = .192$) and a marginal interaction ($p = .052$). In short, in the norm violation condition, the varied agent's actions were seen as much more wrong, and the same was true when the outcome was bad. The marginal interaction suggests that the effect of our manipulation of the varied agent's action might be stronger when the outcome is bad. For ratings of the fixed agent's action valence, there were again main effects of the varied agent's action valence ($p < .007$, $\eta^2_p = .061$) and outcome valence ($p < .001$, $\eta^2_p = .106$) but no interaction ($p = .798$). The fixed agent's actions were seen as worse when the outcome was bad and when the varied agent's actions were neutral.

The ratings of outcome valence are particularly relevant. In contrast to Experiment 1, there was no effect of the varied agent's action on outcome valence, $F(1,116) = .093$, $p = .761$. However, there was a very strong effect of our outcome valence manipulation, with very high ratings for the good outcome ($M = 6.26$, $SD = 1.085$) and very low ratings for the bad outcome ($M = 1.77$, $SD = 1.260$), $F(1,116) = 425.203$, $p < .001$, $\eta^2_p = .786$.

To further verify that judged outcome valence did not account for the causal superseding effect, we re-analyzed causal agreement ratings for the fixed agent in a regression using the varied agent's action as one factor and participants' ratings of outcome valence as another, as well as the interaction term for the two factors. The overall regression was significant, adjusted $R^2 = .250$, $F(3,116) = 14.25$, $p < .001$. Outcome valence ratings were a significant predictor, $\beta = .222$, $p = .006$, as was the varied agent's action, $\beta = -.466$, $p < .001$, but importantly the interaction term was not significant, $\beta = .090$, $p = .258$. Thus, while judged outcome valence did have an independent effect on judgments of the fixed agent's causality, it did not alter the causal superseding effect.

As a final verification that the causal superseding effect exists outside the bounds of motivational accounts, we examined ratings of the fixed agent's causality, but only those in the "good" outcome condition. As the ratings of outcome valence showed, participants regarded this outcome as strongly positive. An excuse validation account would not predict a superseding effect in this case, but there very much is. Even when the outcome is good, participants gave lower agreement ratings for the fixed agent's causality when the varied agent violated a norm ($M = 2.90$, $SD = 2.181$) than not ($M = 4.27$, $SD = 1.760$), $t(59) = -2.681$, $p = .009$, $d = .691$.

3.3. Discussion

Experiment 2 replicated the causal superseding effect and demonstrated that it is not dependent on the valence of the outcome. While outcome valence did have some impact on the causal ratings of the fixed agent, it did not impact the causal superseding effect, that is, the effect of the moral status of the varied agent's actions on the fixed agent's causality. This provides strong evidence that causal superseding can be distinguished from excuse validation (Turri & Blouw, 2014), and therefore goes beyond the predictions of a motivational account. It is particularly striking that the superseding effect emerges even in cases where participants regarded the outcome as positive. Intuitively, one might expect that participants would want to give more credit for a positive outcome to an agent that acted in accordance with a norm, but in fact we find the opposite.

4. Experiment 3

As we have distinguished the causal superseding effect from similar motivational effects, we now turn to two predictions that are wholly unique to the counterfactual account. First, according to the counterfactual account, A will supersede B only if A's action makes B's sufficiency more sensitive. However, in situations where B's sufficiency is robust no matter what A does there should be no causal superseding.

Consider a concrete example: Billy and Suzy work together in the same office. Suzy is supposed to come in at 9 AM, whereas Billy has specifically been told not to come in at that time. The office has a motion detector, and the motion detector will be set off if it detects *two or more people* entering the room at the same time. Both Suzy and Billy arrive at 9 am the next day, and the motion detector goes off. This case has the same basic structure as the ones examined in Experiments 1 and 2, and the counterfactual account predicts that it should produce the same causal superseding effect. Since Billy's action violates a norm, the possibility in which he doesn't act will be seen as highly relevant. Then, since Suzy's act would not be sufficient for the outcome in that possibility, she will be seen as less causal.

But now consider a slightly modified version of the case. What if the motion detector will be set off if it detects *one or more people* entering the room instead? In this case, either Suzy or Billy would have been sufficient to bring about the outcome. Since Billy's action is bad, the possibility in which he doesn't act is seen as highly relevant. However, even in that possibility, Suzy's action would still have been sufficient for bringing about the outcome. Thus, we predict that Suzy's causality should be unaffected by the moral valence of Billy's actions when each individual action is sufficient for bringing about the outcome.

The difference between these two scenarios comes down to a difference in their causal structures. In the first case, and in all of the vignettes used in Experiments 1 and 2, the scenario is *conjunctive*, as the outcome requires the actions of both one agent AND the other. In the second case, where we do not predict causal superseding, the scenario is *disjunctive*, that is, the outcome can be generated by one agent OR the other.

More abstractly, if the varied agent's actions are morally wrong, the possibility that the varied agent does not act becomes more relevant. However, if in that possibility the fixed agent can still bring about the outcome on her own, then her sufficiency is unaffected, and according to the counterfactual sufficiency account, she should not be superseded.

We tested this prediction directly in Experiment 3 by manipulating the event structure such that the scenario was either disjunctive or conjunctive. We predicted a causal superseding effect in the conjunctive scenario but not in the disjunctive scenario.

4.1. Methods

4.1.1. Participants

240 participants were recruited via Amazon Mechanical Turk and paid $0.20 each for completing a brief survey.

4.1.2. Materials and procedure

We created new vignettes that we manipulated along two dimensions. First, as in previous experiments, we manipulated the moral valence of the varied agent's actions, such that they were either neutral or wrong. Second, we manipulated the structure of the event such that both the fixed and varied agent's actions were required to bring about the outcome (conjunctive), or either agent alone could bring about the outcome (disjunctive). See Table 3 for full vignettes.

In all conditions, participants were asked how much they agreed with the statement "Suzy caused the motion detector to go off", using the same 1–7 scale as in previous experiments. (We did not ask about the varied agent's causality in this experiment.) Following this, they were asked to complete a comprehension check: "Who was supposed to show up at 9 am?" They could choose "Billy", "Suzy", or "Both of them."

4.2. Results

We excluded nine participants who failed the comprehension check, leaving 234 for analysis. Fig. 4a shows participants' mean agreement ratings as a function of the moral valence of the varied agent's action and the causal structure of the situation.

A 2 (moral valence) × 2 (causal structure) ANOVA revealed main effects of moral valence, $F(1,230) = 14.666$, $p < .001$, $\eta^2_p = .06$, and causal structure, $F(1,230) = 31.768$, $p < .001$, $\eta^2_p = .121$, as well as a significant interaction between the two, $F(1,230) = 11.577$, $p = .001$, $\eta^2_p = .048$. Further analyses looked at the conjunctive and disjunctive

Table 3
Vignettes for Experiment 3.

(1a) Morally good: 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 9 am. It is absolutely essential that you arrive at that time'.	*(1b) Morally bad:* 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 9 am. 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 9 am.	
(3a) Conjunctive: As it happens, there was a motion detector installed in the room where they arrived. The motion detector was setup to be triggered if *more than one person* appeared in the room at the same time. So the motion detector went off.	*(3b) Disjunctive:* As it happens, there was a motion detector installed in the room where they arrived. The motion detector was setup to be triggered if *at least one person* appeared in the room. So the motion detector went off.

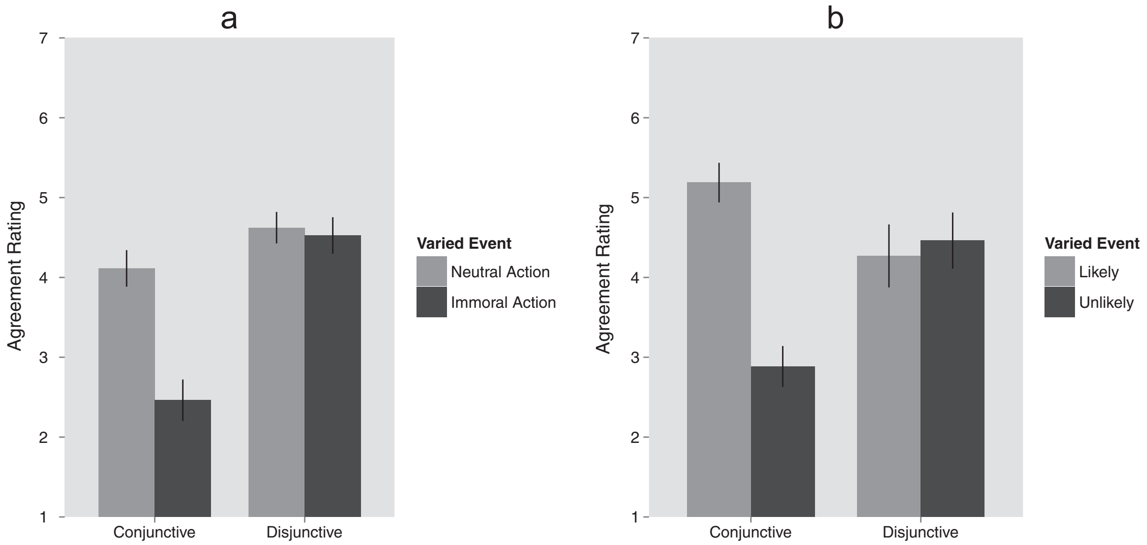

Fig. 4. (a and b): Mean agreement ratings with the causal statements about the fixed agent as a function of causal structure and action valence (Experiment 3, left side) or event probability (Experiment 4, right side). Error bars depict SE mean.

structures separately. As predicted, there was a significant superseding effect in the conjunctive condition, with lower agreement ratings for the fixed agent when the varied agent's actions were morally wrong ($M = 2.46$, $SD = 1.87$) than when they were not ($M = 4.11$, $SD = 1.803$), $t(112) = 4.786$, $p < .001$, $d = .898$. However, in the disjunctive condition, there was no such superseding effect: Agreement ratings for the fixed agent did not differ between situations in which the varied agent's actions were immoral ($M = 4.53$, $SD = 1.76$) or neutral ($M = 4.62$, $SD = 1.54$), $t(118) = .324$, $p = .7$.

These results support the predictions of the counterfactual sufficiency account of causal superseding: Causal superseding occurs only when the actions of one agent can affect the sufficiency of the other agent's action.

5. Experiment 4

In addition to replicating the interaction with causal structure found in Experiment 3, Experiment 4 tested another prediction of the counterfactual sufficiency account. As discussed in the introduction, moral valence is just one example of a violation of norms. Any violation of norms, even non-moral ones, by the varied agent should make the counterfactual possibility that those actions did not occur more relevant. Thus, according to the counterfactual sufficiency account, we should also see causal superseding even when an event is seen as violating a purely statistical norm. Experiment 4 tested this prediction.

5.1. Methods

5.1.1. Participants

120 participants were recruited via Amazon Mechanical Turk and paid $0.20 for their participation.

5.1.2. Materials and procedure

Experiment 4 followed the structure of Experiment 3 very closely, but differed in content. Instead of fixed and varied agents, we used fixed and varied events that resulted from a single agent's actions. The fixed event was a coin-flip, while the varied event was rolling two six-sided dice. We manipulated the likelihood of the varied event by changing the minimum value that the dice needed to achieve in order for the outcome to be successful – higher than 2 (very likely) or higher than 11 (very unlikely). We also manipulated the event structure such that both the coin flip and the die roll were necessary for Alex to win (conjunctive) or either one alone was sufficient (disjunctive). The vignettes for Experiment 4 are displayed in Table 4.

Participants were then asked how much they agreed with the statement, "Alex won because of the coin flip", on a 1–7 scale. They were additionally asked two comprehension check questions: "What did Alex need to roll higher than in order to win?" and "Which was more likely, that he would get heads on the coin flip or roll high enough on the dice roll?"

5.2. Results

13 participants were excluded for having failed to correctly answer the comprehension questions, leaving 107 for analysis. The results can be found in Fig. 4b.

We conducted a 2 (likelihood) × 2 (causal structure) ANOVA. There was a main effect of likelihood, $F(1,106) = 11.294$, $p = .001$, $\eta^2_p = .096$, no main effect of causal structure, $F(1,106) = 1.100$, $p = .297$, but critically, there was once again an interaction between the two, $F(1,106) = 15.786$, $p < .001$, $\eta^2_p = .130$. As in Experiment 3, further analyses revealed that there was a superseding effect only in the conjunctive scenario. In the conjunctive condition, the coin flip was seen as less causal when the

Table 4
Vignettes for Experiment 4.

(1) *Background:* Alex is playing a board game. Every turn in the game, you simultaneously roll two six-sided dice and flip a coin. Alex will either win or lose the game on his next turn.	
(2a) *Likely/conjunctive:* Alex will only win the game if the total of his dice roll is greater than 2 AND the coin comes up heads. It is very likely that he will roll higher than 2, and the coin has equal odds of coming up heads or tails. Alex flips the coin and rolls his dice at exactly the same time. The coin comes up heads, and he rolls a 12, so just as expected, he rolled greater than 2. Alex wins the game.	(2b) *Unlikely/conjunctive:* Alex will only win the game if the total of his dice roll is greater than 11 AND the coin comes up heads. It is very unlikely that he will roll higher than 11, but the coin has equal odds of coming up heads or tails. Alex flips the coin and rolls his dice at exactly the same time. The coin comes up heads, and he rolls a 12, so amazingly, he rolled greater than 11. Alex wins the game.
(2c) *Likely/disjunctive:* Alex will only win the game if the total of his dice roll is greater than 2 OR the coin comes up heads. It is very likely that he will roll higher than 2, and the coin has equal odds of coming up heads or tails. Alex rolls his dice and flips the coin at exactly the same time. The coin comes up heads, and he rolls a 12, so just as expected, he rolled greater than 2. Alex wins the game.	(2d) *Unlikely/disjunctive:* Alex will only win the game if the total of his dice roll is greater than 11 OR the coin comes up heads. It is very unlikely that he will roll higher than 11, but the coin has equal odds of coming up heads or tails. Alex flips the coin and rolls his dice at exactly the same time. The coin comes up heads, and he rolls a 12, so amazingly, he rolled greater than 11. Alex wins the game.

dice roll was unlikely ($M = 2.88$, $SD = 1.31$) than when it was likely ($M = 5.19$, $SD = 1.40$), $t(56) = 6.415$, $p < .001$, $d = 1.704$. However, in the disjunctive condition, the coin flip was equally causal when the dice roll was unlikely ($M = 4.46$, $SD = 1.79$) and likely ($M = 4.27$, $SD = 2.01$), $t(50) = -.364$, $p = .7$.

6. General discussion

Four experiments demonstrated the phenomenon of causal superseding and found supporting evidence for the predictions of a counterfactual sufficiency account. Experiments 1 and 2 demonstrated that the effect operates outside the bounds of excuse validation and other motivational accounts. Experiments 3 and 4 showed that the effect holds in conjunctive causal structures but not disjunctive causal structures, as predicted by a counterfactual sufficiency account. Finally, Experiment 4 demonstrated that causal superseding is not specific to violations of moral norms, but shows up for violations of statistical norms as well.

The causal superseding effect demonstrated across these four studies is both an exciting discovery and something that has been in legal records for over a century (Hart & Honoré, 1985). The case of *Carter v. Towne* and other legal decisions show that this effect emerges in real-world contexts and can have a large impact on the lives of those involved in these court decisions. At the same time we have demonstrated something very surprising in the context of previous work on causal reasoning: Causal judgments of an agent's role in neutral or even positive outcomes (such as deleting computer viruses, or winning a board game) can be strongly affected not only by their own actions, but by the actions of other agents, provided the event has a particular causal structure.

At this point, it is important to acknowledge two relatively recent findings that provide evidence for effects that are related to, but distinct from, causal superseding. McGill and Tenbrunsel (2000) examined causal judgments in cases in which two causal factors combined in a conjunctive way to bring about an outcome. Their experiments varied the ease with which one of the causal factors could be imagined to have been different (mutability) and the likelihood that this factor will bring about the outcome (propensity). The results showed that varying the mutability and propensity of one causal factor influenced participants' causal ratings judgments about the other factor. More specifically, McGill and Tenbrunsel (2000) found an interaction between mutability and propensity, such that propensity had the opposite effect depending on whether the cause was more or less mutable. When the alternative causal factor had a high propensity to bring about the effect, causal judgments to the target factor decreased when the alternative factor's mutability was high compared to low. This pattern of results is consistent with the predictions of the counterfactual sufficiency model. In contrast, when the alternative causal factors' propensity was low, the target factor was seen as more causal when the alternative factor's mutability was high rather than low. In our scenarios, we did not manipulate mutability and propensity at the same time. More work is needed to explore the ways in which variations in mutability, propensity, and causal structure influence causal judgments.

A second related study made a valuable contribution to the current work already, in providing a framework for the vignettes in Experiment 2 (Reuter et al., 2014). In the original study, they examined the role of norm violations and temporal order in causal selection. Importantly, rather than using a rating scale, Reuter et al. (2014) had their participants make a forced choice between the two agents in the scenario. They found that the agent who violated a norm was more likely to be selected as the cause, and indeed this is compatible with the causal superseding effect, but it does not distinguish whether the varied actor was seen as more causal or whether the fixed actor was seen as less causal. Temporal order also played a significant role in their findings, but was not a focus of the current research. However, the roles of temporal order and contingency are critical avenues of future research.

In the remainder of our discussion, we will delve further into our counterfactual account and present a number of specific and testable additional predictions as avenues for possible future research. Regardless of whether or not they turn out to be precisely correct, the phenomenon of causal superseding is worthy of study, and testing possible explanations should grant novel insight into causal and counterfactual reasoning.

6.1. Normality and counterfactuals

One key element of our account is the notion that moral judgments impact causal cognition by playing a role in people's overall judgments of *normality* (Hitchcock & Knobe, 2009). Hence, we predicted, and found, that non-moral norm violations have the same effect as moral violations. Our account merely requires a norm violation that leads people to focus on particular counterfactual possibilities, but makes no stipulations about the nature of the norm being violated. Thus, we expected that any norm violation should yield a causal superseding effect.

The strongest support for this can be found in comparing the results of Experiments 3 and 4. Experiment 3 used a moral norm violation, whereas Experiment 4 used a statistical norm violation. The two experiments were otherwise extremely similar, and the results, while differing slightly in the strength of the effect, show the same clear pattern. Rather than just a simple main effect, both experiments produced the same interaction between causal structure and norm violation. Someone who wished to argue that we must treat these norm violations differently, or that we found two different superseding effects, would face a steep challenge.

It is worth noting that the counterfactual possibilities that people consider based on these norm violations is partially reliant on the background knowledge they bring to the scenario. In using a variety of scenarios designed to be close to the real world, we relied on participants sharing certain assumptions about the real world such that they would consider the right counterfactual possibilities, and consider our intended norm violations to actually be norm violations. For example, one crucial assumption of the Bartlett Bookends scenario is the assumption that, had Bill not stolen the bookend from his friend, he would have had no other means of acquiring a right-side Bartlett bookend. We emphasize this to participants by stating that his friend could not bear to part with it, and failing to mention a right-side Bartlett bookend anywhere else in the scenario, but ultimately it is participants' own knowledge about how the world works that they are bringing to the experiment that determines which counterfactual possibilities they consider.

However, in the context of our explanation, norm violations are merely one way in which particular counterfactual possibilities are highlighted. In fact, our explanation does not require norm violations at all. The causal superseding effect merely requires a salient counterfactual possibility in which the sufficiency of one cause is undermined. Any other means of making such a possibility salient should generate the superseding effect as well. For example, one might be able to generate a superseding effect by explicitly instructing participants to consider a specific counterfactual possibility. Or one could make use of any of the other factors that have been studied in regard to counterfactual thinking such as the controllability of the action or outcome (c.f., Girotto, Legrenzi, & Rizzo, 1991; McCloy & Byrne, 2000; McGill & Tenbrunsel, 2000).

6.2. Sufficiency and sensitivity

Our explanation of the causal superseding effect rests on two key assumptions. First, people's causal judgments are influenced by the degree to which they regard a factor as sufficient and insensitive. In other words, when people are trying to determine whether A caused B, their judgments are influenced in part by the degree to which they think A would have been sufficient for B in various counterfactual possibilities. Second, people do not treat all counterfactual possibilities equally. They regard some counterfactual possibilities as more relevant than others. Thus far, we have been relying on a purely computational-level understanding of these two assumptions. We now present the broad outlines of an approach to actually describing them on a more algorithmic level.

Our approach takes advantage of an insight that has proven helpful in numerous other areas of cognitive science (e.g., Denison, Bonawitz, Gopnik, & Griffiths, 2013). Specifically, we propose that people might solve this problem by *sampling*. In other words, it is not that people consider every single counterfactual possibility and then weight each possibility by its degree of normality. Rather, people simply sample a small number of counterfactual possibilities, with the probability of any given counterfactual possibility ending up in the sample being proportional to its degree of normality.

To present this approach in more formal terms, we turn to Causal Bayes Nets (Halpern & Pearl, 2005; Pearl, 2000), in which causal relationships are defined in terms of functional relationships between variables representing potential causes and effects. Causal Bayes Nets take the form of networks of variables and causal relationships between them that represent the probability distribution for one set of variables given that one observes or sets values for another set of variables. Therefore they can support inferences about the state of causal variables from observed effects, and are useful for predicting the outcomes of interventions on specific variables. Causal Bayes Nets can also be used to represent counterfactual statements about what would have happened if the state of a particular variable had been different from what it actually was. The counterfactual aspects of Causal Bayes Nets will be the focus of our discussion here.

With Causal Bayes Net representations of counterfactual possibilities in the background, let us now consider Pearl's (1999) definition of counterfactual sufficiency. To assess whether or not a candidate causal variable was sufficient for bringing about the effect, we must first condition on what actually happened. Now, we imagine a situation in which the candidate causal event as well as the effect event did *not* occur. In this possibility in which the effect event did not occur, we then assess whether intervening on the candidate cause in order to make it true would reestablish the effect we have actually observed. If the effect would still have occurred following that intervention, then the cause was sufficient for bringing it about; otherwise, it was not.

Intuitively, we can think of this operation in the following way: in order to assess the counterfactual sufficiency of a causal event, we undo the specific events leading up to the causal event of interest, effectively "rewinding" the world to the point before the event occurs, though leaving factors not on the direct pathway from (proposed) cause to effect as they are. In that minimally different situation, we then imagine that the causal event was in fact true, "press the play button" in the mental simulation, and see whether or not the effect would still have occurred.

Recently, Lucas and Kemp (2012) have extended Pearl's definition in a way that introduces the notion of sampling. Rather than supposing that people use a veridical, "rewound" copy of the actual world, they suggest that people 'resample', yielding a noisy copy of the actual world. Thus, people do not simply get a purely deterministic answer as to whether the outcome would have still occurred, but an answer that draws on a probabilistic sampling process. Yet this hypothesis immediately leaves us with a further question: Namely, how to model the process whereby people sample possibilities.

It is here that the notion of *norms* become relevant to our account. As we saw above, people are more inclined to consider possibilities that unfold in accordance with norms (both statistical and moral) than to consider possibilities that violate such norms. Thus, if a variable realizes a norm-violating value in the actual world (Bill steals a bookend), its value in the simulation, if resampled, is more likely to be in accordance with the norm (Bill does not steal the bookend).

Putting these ideas together, we get at least the broad outlines of how causal superseding can be explained through the influence of normality on a model of counterfactual sufficiency. The basic idea is that people are implicitly running mental simulations over their causal representation of the world. To determine whether A is sufficient for B, they run a series of simulations of a certain type in which they set A to occur and check to see whether B occurs. However, they do not simply run simulations that differ from the actual world in arbitrary ways. Instead, they show a bias to simulate possibilities in which events accord with norms rather than violate norms. For this reason, if A only brings about B in conjunction with some norm-violating event, A will be unlikely to be seen as sufficient for B. This is the phenomenon of causal superseding.

6.3. Integrating sufficiency into a larger theory

In the previous section, we have presented a theory of counterfactual sufficiency that accounts well for the results of the experiments reported in this paper. While we do believe that causal judgments are closely linked to counterfactual simulations over causal representations (cf. Chater & Oaksford, 2013; Gerstenberg et al., 2014; Goodman, Tenenbaum, & Gerstenberg, in press), we have also made it clear that the proposed account of the superseding effect is not to be understood as a complete theory of causal attribution. Indeed, there are several empirical phenomena that our counterfactual sufficiency account would not apply to.

First, while our account of the superseding effect focuses on counterfactual sufficiency, several studies have shown that people's causal judgments are influenced by counterfactual necessity as well. For example, previous work has shown that people's causal and responsibility attributions are reduced in situations in which the outcome was causally overdetermined by multiple individually sufficient causes (Gerstenberg & Lagnado, 2010, 2012; Lagnado et al., 2013; Zultan et al., 2012). When two players in a team succeeded in their individual task, each player received greater responsibility for the team's win in a situation in which both contributions were necessary compared to a situation in which the success of either player in the team would have been sufficient. This effect cannot be explained in terms of counterfactual sufficiency. The sensitivity of whether an agent's action is sufficient for the outcome is increased in a conjunctive causal structure but is unaffected by the (expected) actions of others in a disjunctive causal structure.

Building on formal structural models of causal responsibility (Chockler & Halpern, 2004; Halpern & Pearl, 2005), Lagnado et al. (2013) developed a *criticality-pivotality model* which predicts that people's responsibility attributions to individual group members are influenced both by (a) how critical each individual's contribution is perceived for the group's positive outcome ex ante and by (b) how close each individual's contribution was to being pivotal ex post (see Lagnado & Gerstenberg, in press, for evidence that similar considerations also influence people's judgments in non-agentive contexts). Note that the pattern of causal judgments found in the current paper would not be predicted by the criticality-pivotality model. The criticality-pivotality model predicts that the extent to which a person is judged to be causally responsible for an outcome decreases the more distant the actual situation was from a situation in which the person's action would have made a difference to the outcome. However, the experiments reported here show that the causal superseding effect occurs in conjunctive situations (where each person's contribution *was* pivotal) but not in disjunctive situations (where the outcome was overdetermined and hence neither contribution pivotal). Future research will need to tease apart the factors that determine in what situations causal judgments are more strongly influenced by counterfactual necessity versus sufficiency.

Second, counterfactual sufficiency does not explain the effect whereby abnormal actions are judged more causal than normal actions. It does predict that the normality of A's actions influences the causal judgment about B's actions when their actions combine conjunctively – that is the basic superseding effect. However, in line with previous findings (Alicke, 1992; Hitchcock & Knobe, 2009), our results show that A is judged more causal when her action was norm-violating than when it conformed to the norm. Thus, as we noted in the introduction, judgments of sufficiency appear to be one part of the story but do not by any means constitute the entirety of causal cognition.

One way to incorporate previous findings into our framework would be to assume that the normality of A's action influences whether people consider a counterfactual possibility in which A behaves differently. When A's action was abnormal, they might consider whether A's action would still have been sufficient for bringing about the outcome if it had been normal instead (cf. Halpern & Hitchcock, 2014; Hitchcock & Knobe, 2009). However, when A's action was normal, counterfactual possibilities involving an alternative abnormal action do not come to mind so naturally. Thus, A's action is perceived to have made more of a difference to the outcome when it was abnormal than when it was normal (Petrocelli et al., 2011).

6.4. Broader implications for causal cognition

One point that has not yet been discussed is how we treat the role of moral and other norm violations in causal cognition. Previous work has been divided on whether we should treat moral considerations as (a) playing a role in the operation of people's causal cognition itself or (b) introducing some external bias or pragmatic factor that is skewing the results of what is in fact a purely non-moral causal cognition system (for a review, see Knobe, 2010). We present further evidence that moral violations are included in our causal reasoning, in part by demonstrating an effect that, as noted above, does not hinge on morality specifically. The causal superseding effect has two key features which make it difficult to explain as a bias or pragmatic effect that is not part of causal cognition: First, we find it with non-moral norm violations, something which these morality-as-external-bias accounts do not suggest. The probabilistic norm violation in Experiment 4 provides evidence that norm violations in general, including moral violations, have an impact on our causal judgments. Second, the effect of both moral and statistical norm violations is dependent on the causal structure of the event. If morality functioned as an outside bias that skewed our causal judgment, it would be somewhat surprising if that bias operated on only some causal structures and not others. Experiments 3 and 4 suggest that these norm violations are considered along with causal structure in making causal judgments.

6.5. Conclusion

This paper presented evidence for the surprising phenomenon of causal superseding, and offered a preliminary explanation that opens up a number of avenues for further exploration. Beyond the effect of causal superseding itself, this paper brings to light a number of interesting questions about how the actions of one agent can impact causal judgments about other agents. These effects are worth investigating both for the real-world impact these judgments can have in legal settings and for the insights they can afford us into the operations of human causal cognition.

Acknowledgements

TG was supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. DL was supported by ESRC Grant RES-062330004. The authors are deeply grateful to Scott Shapiro for his contributions to the early stages of the project.

References

Alicke, M. D. (1992). Culpable causation. *Journal of Personality and Social Psychology, 63*(3), 368.

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin, 126*(4), 556–574.

Alicke, M. D., Buckingham, J., Zell, E., & Davis, T. (2008). Culpable control and counterfactual reasoning in the psychology of blame. *Personality & Social Psychology Bulletin, 34*(10), 1371–1381. http://dx.doi.org/10.1177/0146167208321594.

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

Byrne, R. M. (2005). *The rational imagination: How people create alternatives to reality*. The MIT Press.

Chater, N., & Oaksford, M. (2013). Programs as causal models: Speculations on mental programs and mental representation. *Cognitive Science, 37*(6), 1171–1191.

Chockler, H., & Halpern, J. Y. (2004). Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research, 22*(1), 93–115.

Denison, S., Bonawitz, E., Gopnik, A., & Griffiths, T. L. (2013). Rational variability in children's causal inferences: The sampling hypothesis. *Cognition, 126*(2), 285–300. http://dx.doi.org/10.1016/j.cognition.2012.10.010.

Fincham, F. D., & Roberts, C. (1985). Intervening causation and the mitigation of responsibility for harm doing II. The role of limited mental capacities. *Journal of Experimental Social Psychology, 21*(2), 178–194.

Fincham, F. D., & Shultz, T. R. (1981). Intervening causation and the mitigation of responsibility for harm. *British Journal of Social Psychology, 20*(2), 113–120. http://dx.doi.org/10.1111/j.2044-8309.1981.tb00483.x.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2014). From counterfactual simulation to causal judgment. In *Proceedings of the 36th annual conference of the cognitive science society*, Austin, TX, 2014. Cognitive Science Society.

Gerstenberg, T., & Lagnado, D. A. (2010). Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition, 115*(1), 166–171.

Gerstenberg, T., & Lagnado, D. A. (2012). When contributions make a difference: Explaining order effects in responsibility attributions. *Psychonomic Bulletin & Review, 19*(4), 729–736.

Girotto, V., Legrenzi, P., & Rizzo, A. (1991). Event controllability in counterfactual thinking. *Acta Psychologica, 78*(1), 111–133. http://dx.doi.org/10.1016/0001-6918(91)90007-M.

Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (in press). Concepts in a probabilistic language of thought. In Margolis & Lawrence (Eds.), *The conceptual mind: New directions in the study of concepts*. MIT Press.

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

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

Hart, H. L. A., & Honoré, T. (1985). *Causation in the law*. Oxford: Oxford University Press.

Hitchcock, C. (2012). Portable causal dependence: A tale of consilience. *Philosophy of Science, 79*(5), 942–951.

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

Kahneman, D., & Miller, D. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 80*, 136–153.

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences, 33*(4), 315–365.

Knobe, J., & Szabó, Z. G. (2013). Modals with a taste of the deontic. *Semantics and Pragmatics, 6*, 1–42.

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2014). Causal supersession. In *Proceedings of the cognitive science society 2014 annual meeting*. <https://mindmodeling.org/cogsci2014/>.

Lagnado, D. A. & Gerstenberg, T. (in press). A difference-making framework for intuitive judgments of responsibility. In D. Shoemaker (Ed.), *Oxford studies in agency and responsibility*. Oxford University Press.

Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition, 108*(3), 754–770. http://dx.doi.org/10.1016/j.cognition.2008.06.009.

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science, 47*, 1036–1073.

Lewis, D. (1973). Causation. *The Journal of Philosophy, 70*(17), 556. http://dx.doi.org/10.2307/2025310.

Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology, 61*, 303–332.

Lucas, C. G., & Kemp, C. (2012). A unified theory of counterfactual reasoning. In *Proceedings of the 34th annual conference of the cognitive science society*.

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

McCloy, R., & Byrne, R. M. (2000). Counterfactual thinking about controllable events. *Memory & Cognition, 28*(6), 1071–1078.

McClure, J., Hilton, D. J., & Sutton, R. M. (2007). Judgments of voluntary and physical causes in causal chains: Probabilistic and social functionalist criteria for attributions. *European Journal of Social Psychology, 37*(5), 879–901.

McGill, A. L., & Tenbrunsel, A. E. (2000). Mutability and propensity in causal selection. *Journal of Personality and Social Psychology, 79*(5), 677. http://dx.doi.org/10.1037/0022-3514.79.5.677.

N'gbala, A., & Branscombe, N. R. (1995). Mental simulation and causal attribution: When simulating an event does not affect fault assignment. *Journal of Experimental Social Psychology, 31*(2), 139–162.

Pearl, J. (1999). Probabilities of causation: Three counterfactual interpretations and their identification. *Synthese, 121*(1–2), 93–149.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge University Press.

Petrocelli, J. V., Percy, E. J., Sherman, S. J., & Tormala, Z. L. (2011). Counterfactual potency. *Journal of Personality and Social Psychology, 100*(1), 30–46. http://dx.doi.org/10.1037/a0021523.

Reuter, K., Kirfel, L., van Riel, R., & Barlassina, L. (2014). The good, the bad, and the timely: How temporal order and moral judgment influence causal selection. *Frontiers in Psychology, 5*, 1336. http://dx.doi.org/10.3389/fpsyg.2014.01336.

Spellman, B. A. (1997). Crediting causality. *Journal of Experimental Psychology: General, 126*(4), 323–348.

Teigen, K. H., & Brun, W. (2011). Responsibility is divisible by two, but not by three or four: Judgments of responsibility in dyads and groups. *Social Cognition, 29*(1), 15–42.

Turri, J., & Blouw, P. (2014). Excuse validation: A study in rule-breaking. *Philosophical Studies, 172*(3), 615–634. http://dx.doi.org/10.1007/s11098-014-0322-z.

Wells, G. L., & Gavanski, I. (1989). Mental simulation of causality. *Journal of Personality and Social Psychology, 56*(2), 161–169.

Woodward, J. (2006). Sensitive and insensitive causation. *The Philosophical Review, 115*(1), 1–50.

Zultan, R., Gerstenberg, T., & Lagnado, D. A. (2012). Finding fault: Counterfactuals and causality in group attributions. *Cognition, 125*(3), 429–440.