Word Count (including references and notes): 8,734

Causal Attributions and the Trolley Problem

Justin Sytsma
Victoria University of Wellington, Wellington, NZ

Jonathan Livengood
University of Illinois, Urbana-Champaign

Abstract. In this paper, we consider two competing explanations of the empirical finding that people's causal attributions are responsive to normative details, such as whether an agent's action violated an injunctive norm—the *counterfactual view* and the *responsibility view*. We then present experimental evidence that uses the trolley dilemma in a new way to investigate causal attribution. In the switch version of the trolley problem, people judge that the agent ought to flip the switch, but they also judge that she is more responsible for the resulting outcome when she does so than when she refrains. As predicted by the responsibility view, but not the counterfactual view, people are more likely to say that the agent caused the outcome when she flips the switch.

Keywords. causal attributions, actual causation, responsibility, counterfactual, trolley problem

Causal Attributions and the Trolley Problem

Metaphysicians working on causation face a dilemma. On the one hand, injunctive norms are supposed to be irrelevant to what causes what.[1] As Helen Beebee (2004, 293) puts it, "no philosopher working within the tradition I'm concerned with here thinks that the *truth* conditions for causal claims contain a moral element." On the other hand, many philosophers working in the tradition Beebee references have endorsed the claim that accounts of the metaphysics of causation should be informed or constrained by ordinary intuitions, ordinary attributions, ordinary concepts, common sense, or the like (e.g., Lewis, 1986; Menzies, 1996; Collins, Hall, and Paul, 2004; Hall, 2004; Liebesman, 2011; Paul and Hall, 2013; Halpern and Hitchcock, 2015; Halpern, 2016), and a large body of research indicates that injunctive norms matter for ordinary causal attributions.[2]

[1] Injunctive norms include both prescriptive norms (what people *should* do) and proscriptive norms (what people should *not* do). In the existing literature, authors often use the expression "prescriptive norm" to refer to both prescriptive *and* proscriptive norms. We find that usage infelicitous, so we adopt the term "injunctive norm" instead. Injunctive norms can be distinguished from descriptive norms (often referred to as "statistical norms"). Some philosophers (e.g., Hitchcock and Knobe, 2009) accept a further category of norms of proper functioning, which apply to designed systems, including systems "designed" by natural selection. There is ongoing debate about whether descriptive norms have an independent effect on ordinary causal attributions or whether they simply play a role in mediating injunctive norms (e.g., Knobe and Fraser 2008; Sytsma, Livengood, and Rose 2012; Livenood, Sytsma, and Rose 2017). When we talk about norms in this paper, we'll be leaving descriptive norms to the side, focusing on situations in which people judge that an agent ought, or ought not, perform a particular action. Further, while some work has been done on cases involving positive outcomes (e.g., Alicke, Rose, and Bloom 2011), we'll focus on cases involving negative outcomes. More generally, philosophers and psychologists have identified many interesting aspects of the process of causal attribution that we will not touch on in this paper. See Livengood and Rose (2016), Halpern and Hitchcock (2015), Kominsky et al. (2015), and Halpern (2016) for much more detail.

[2] Philosophers working on causation today are divided with respect to the proper target of their inquiry. On the one hand, there are philosophers who we might describe as *conceptualists*. Conceptualists think that we should either offer an analysis of the ordinary concept of causation or offer a suitably refined explication of it. On the other hand, there are philosophers who we might describe as *realists*. Realists think that we should identify truths about the causal relation, just as scientists have identified truths about planets and electrons. While the relevance of empirical work is probably most obvious in connection with conceptualist approaches, we believe that philosophers of both conceptualist and realist persuasions should be interested in ordinary causal attributions, though possibly for very different reasons. We will not explore possible bridge principles from facts about ordinary causal attributions to philosophical theories of causation in this paper. For discussion of some general strategies, see Sytsma and Livengood 2015.

In case after case, participants assign higher causal ratings to an individual who contributes to bringing about a negative outcome when she violates an injunctive norm than when she does not. This includes both cases where an agent violates an injunctive norm through her actions (e.g., Hilton and Slugoski 1986; Alicke 1992; Knobe and Fraser 2008; Hitchcock and Knobe 2009; Sytsma, Livengood, and Rose 2012; Reuter et al. 2014; Kominsky et al. 2015; Livengood, Sytsma, and Rose 2017) and where she violates an injunctive norm through her inaction (e.g., Sytsma ms). Here is an illustration. We presented participants with the following version of the e-mail case tested by Kominsky et al. (2015)[3], in which we have changed the norm-violating agent's action to an inaction:

> Billy and Suzy work for a company that has a central computer. Both Billy and Suzy answer incoming calls.
>
> Since incoming calls are logged through the central computer, 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 supposed to log into the central computer in the mornings, whereas Billy is supposed to log into the central computer in the afternoons. Further, the policy states that both Billy and Suzy are supposed to log out of the central computer when they are done answering calls.
>
> Unfortunately, a problem has recently developed with the computer system: if two people are logged into the central computer at the same time, some important work e-mails will be immediately deleted.
>
> Yesterday, Billy logged into the central computer in the afternoon. He did not log out when he was done, however, because a pirated movie he wanted to watch was still downloading. This morning, Suzy logged into the central computer to answer incoming calls. Since two people were logged into the central computer at the same time, some important work e-mails were deleted.

[3] A similar case is presented in Knobe (2006), and variations have been tested in Livengood, Sytsma, and Rose (2017) and Reuter et al. (2014).

Participants were then asked how strongly they agreed or disagreed that "Billy caused the e-mails to be deleted" and that "Suzy caused the e-mails to be deleted," answering on a 7-point scale running from "Strongly Disagree" to "Strongly Agree."[4, 5]

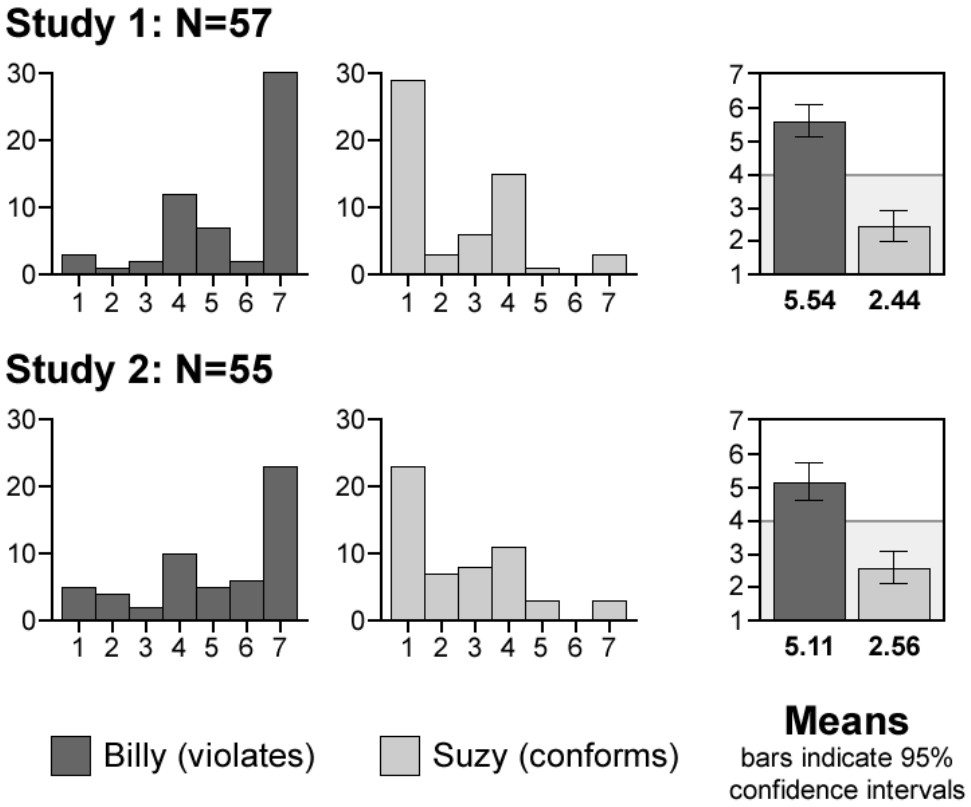

Figure 1: Results of Studies 1 and 2

In line with previous results, we found that people tended to agree that the norm-violating agent (Billy) caused the e-mails to be deleted (mean=5.54 with a 95% CI of [5.06, 6.02]), but to disagree that the norm-conforming agent (Suzy) caused the e-mails to be deleted (mean=2.44

[4] In all studies reported in this paper, responses were collected online through advertising on Google Ads; participants were native English speakers, 16 years of age or older, with at most minimal training in philosophy. Minimal training in philosophy was taken to exclude philosophy majors, those who have completed a degree with a major in philosophy, and those who have taken graduate-level courses in philosophy.

[5] Responses were collected from 57 participants—68.4% women, with an average age of 32.8 years, and ranging in age from 16 to 79.

with a 95% CI of [1.98, 2.90]).[6] Similar results were found in a second study where we made both the norm-violating and the norm-conforming characters inactive. Participants were again statistically more likely to answer that the norm-violating agent caused the problem (mean=5.11 with a 95% CI of [4.55, 5.67]) than that the norm-conforming agent caused the problem (mean=2.56 with a 95% CI of [2.10, 3.03]), even though neither took any positive action.[7] The results of these studies are shown in Figure 1.

Taken at face value, these and related empirical findings suggest that according to the ordinary conception of causation, the truth conditions for causal claims contain a normative element.[8] But if this is correct, then either we were wrong to think that injunctive norms are irrelevant to a properly metaphysical account of what causes what or we were wrong to think that accounts of the metaphysics of causation should be informed or constrained by ordinary intuitions, ordinary attributions, ordinary concepts, common sense, or the like. On our account, which we call the *responsibility view*, the ordinary concept of causation does have normative content. Of course, one need not take the empirical findings at face value. One prominent way out of the dilemma is to endorse a *counterfactual view*, according to which injunctive norms matter to causal attributions because of their effect on our counterfactual judgments but the concept of causation itself has no normative content.

---

[6] Causal ratings for Billy were statistically above the neutral point whether using a t-test (parametric) or Wilcoxon signed rank test (nonparametric): t(56)=6.46, p=2.7e-8; V=927, p=1.2e-6. By contrast, causal ratings for Suzy were statistically below the neutral point for each test: t(56)=-6.81, p=7.3e-9; V=83.5, p=1.1e-6. And the difference between the two was significant with a large effect size: t(111.82)=9.38, p=9.3e-16, Cohen's d=1.76; W=2837.5, p=1.4e-12, Cliff's delta=0.75. The same statistical tests will be used for subsequent analyses. Going forward, we will say that a rating is "statistically" above or below some value rather than using the awkward expression "statistically significantly." When we say that some rating was statistically above or below a specific value, we mean that a test of significance rejected the hypothesis of no difference.

[7] See the supplemental materials for the full text of the probes used in the studies reported in this paper. Responses for Study 2 were collected from 55 participants—76.4% women, with an average age of 29.4 years, and ranging in age from 16 to 72. Causal ratings for Billy were statistically above the neutral point: t(54)=3.97, p=0.00021; V=819.5, p=0.00044. Causal ratings for Suzy were statistically below the neutral point: t(54)=-6.20, p=8.2e-8; V=112.5, p=4.5e-6. And the difference between the two was significant with a large effect size: t(104.47)=7.02, p=2.4e-10, Cohen's d=1.34; W=2461.5, p=7.3e-9, Cliff's delta=0.63.

[8] By "taken at face value," here, we mean roughly that ordinary attributions are accurately reflecting an underlying concept that is being competently deployed.

In this paper, we present new empirical evidence that is predicted by the responsibility view but difficult to square with the counterfactual view. We thereby offer a challenge to metaphysicians who want to avoid the dilemma by way of a counterfactual view.[9] We consider a case in which people judge that an agent ought to act in a way that would leave her responsible for a bad outcome. We find that people's causal attributions are strikingly similar to their responsibility attributions. When an agent is said to be responsible for a bad outcome, she is also said to have caused the outcome, even when she is judged to have acted morally. We argue that while the responsibility view directly explains our findings, the counterfactual view predicts the opposite effect.

Here is how we will proceed. We begin in Section 1 by describing the counterfactual and responsibility views. In Section 2, we detail the test case we will focus on—the switch version of the trolley problem—and lay out the competing predictions that our two target views make about it. We report our experimental results in Section 3. Finally, in Section 4 we critically consider how advocates of counterfactual relevance accounts might respond to our findings.

1. Two Explanations

The counterfactual view holds that norm violations matter for ordinary causal attributions because people are more likely to consider counterfactual alternatives where the norm violation is replaced with a norm-conforming event (e.g., Hitchcock and Knobe 2009, Halpern and Hitchcock 2015, Kominsky et al. 2015, Icard et al. 2017). Perhaps the most widely discussed version of this view is the one put forward by Hitchcock and Knobe (2009). They argue that

[9] There are, of course, other explanations of the effect of norms on ordinary causal attributions that metaphysicians might call on to avoid the dilemma—most notably, Alicke's *bias view* (1992, 2000; Alicke, Rose, and Bloom 2011; Rose 2017) and Samland and Waldman's *pragmatic view* (2015, 2016; Samland et al. 2016). We cannot address every avenue of retreat here. But see Sytsma et al. (2019) and Livengood and Sytsma (forthcoming) for discussion of the bias and pragmatic views.

causal attributions serve to identify suitable intervention points, and norm violations come into play because they affect which counterfactuals are most salient for determining whether or not an intervention point is suitable. The basic idea is that in considering a situation, people think about how the outcome could have been prevented. But they don't consider *just any* way the outcome might have been prevented. Instead, they focus on abnormal aspects of the situation in which the outcome has occurred.[10]

The basic picture, then, is that people make causal attributions through a process of counterfactual-based reasoning guided by the evaluation of norms.[11] People identify the cause(s) of a given effect by testing a collection of counterfactual conditionals where the antecedent says that one of various potential causes (which actually occurred) did not occur and the consequent says that the effect did not occur. People are (implicitly) guided by overall normality judgments in deciding which counterfactuals to test. Counterfactuals in which the antecedent is more normal than the actual state of affairs are checked first or are given greater weight or are regarded as more salient or more relevant. Hence, causal attributions are guided by judgments of overall normality, and the result of this process serves to identify suitable intervention points. As such, norms play a significant role in the generation of causal attributions, but the ordinary concept of causation at play in causal attributions is not taken to be a normative concept. Rather, injunctive norms are part of the background against which causal attributions are made, with injunctive norms coming into play when people attempt to identify intervention points in normatively-laden situations.

Let's illustrate this explanation by applying it to the e-mail case discussed in the introduction. Important e-mails will be deleted if two people are logged into the central system at

[10] Hitchcock and Knobe focus on overall normality judgments, which include descriptive as well as injunctive norms, but in the cases we're concerned with, the dominant norms are injunctive norms.

[11] As Hitchcock and Knobe observe, the thought that norm-violation matters to causal attributions is not new. See for example, Hart and Honoré (1985, Chapter 2, Section IIIA) or Hilton and Slugoski (1986).

the same time. Both Billy and Suzy come to be logged into the system, and the e-mails are consequently deleted. When asked about this scenario, participants overwhelmingly agree that Billy caused the outcome, but disagree that Suzy caused the outcome. Why do we find this asymmetry in causal ratings, given that *both* agents needed to be logged into the system for the e-mails to be deleted? The key difference is that Billy violated company policy, while Suzy followed it. According to the counterfactual view, Billy violating the injunctive norm makes salient the counterfactual alternative on which he follows the policy. If Billy had followed the policy, the e-mails would not have been deleted when Suzy logged into the system. This marks Billy as a clear point for intervention: one obvious way to prevent e-mails from being deleted in the future is to make sure that employees follow the policy.

The counterfactual view treats the ordinary concept of causation at play in causal attributions as descriptive. By contrast, on our responsibility view (Sytsma, Livengood, and Rose 2012; Livengood, Sytsma, and Rose 2017; Sytsma et al. 2019; Livengood and Sytsma forthcoming), the concept is not purely descriptive but has built-in normative content. On our view, causal attributions generally serve to issue normative evaluations. Saying that an agent *caused* an outcome, for instance, typically serves to indicate something more than that the agent *brought about* that outcome or that the agent's action *produced* that outcome. Rather, it serves to express a normative evaluation that can be roughly captured by saying that the agent is *responsible for* that outcome or that the agent is *accountable for* that outcome. Putting this in terms of concepts, rather than the corresponding terms, we hold that the dominant ordinary concept of causation at play in causal attributions is relevantly akin to the ordinary concept of responsibility.

To illustrate, consider the e-mail case again. Billy's inaction and Suzy's action jointly lead to a bad outcome. But while Billy violates company policy, Suzy follows it. If we want to figure out who should be held accountable for the resulting problem, the answer seems clear:

Billy is responsible for the e-mails being deleted, not Suzy. According to the responsibility view, causal attribution is akin to expressing such a judgment of normative responsibility. Billy caused the e-mails to be deleted; Suzy did not. But *why* is Billy responsible for the outcome and not Suzy? Both agents played a role in the e-mails being deleted. If Suzy had not logged in, the problem would not have occurred; and, if Billy had logged out, the problem would not have occurred. So why do we hold Billy alone accountable? Billy acted negligently (violating the policy) while Suzy acted responsibly (following the policy). Further, while the problem could have been prevented by Suzy not logging in, this would have created a new problem for the company in that Suzy wouldn't have been able to answer incoming calls. No such problem would have been created by Billy logging out.

Our explanation of the pattern of findings for the e-mail case makes use of counterfactuals, but unlike the counterfactual view, the responsibility view uses counterfactuals to assess the normative difference between different agents' behaviors. According to the counterfactual view, ordinary causal attributions are asymmetric because people do not consider equally both what would have happened if Billy had logged out and also what would have happened if Suzy had not logged in. As such, counterfactual saliency might be seen as a type of heuristic that leads to systematically mistaken causal attributions (i.e., a bias) in scenarios like the e-mail case.

We do not hold that people's causal attributions about scenarios like the e-mail case are mistaken. We think they simply follow the dominant usage, which has a normative dimension. As such, our explanation does not hinge on participants considering one counterfactual rather than another, and we expect participants would tend to arrive at the same judgments if they considered both counterfactuals noted above. In other words, we do not explain the asymmetry in causal ratings in terms of a bias toward one counterfactual, even though we accept that

counterfactual reasoning will be involved in assessing the relative normative responsibility of each agent for the outcome.

Unfortunately, despite their differences, the counterfactual and responsibility views make the same predictions about a wide range of cases, including many prominent cases from the literature. The reason is that in the most commonly tested scenarios two agents jointly bring about an outcome, while the normative status of the two agents' behaviors are varied. In such cases, the counterfactual view predicts higher causal ratings for norm-violating agents (because counterfactuals on which the agents conform to a norm will be more salient), and the responsibility view makes the same prediction (because of the normative difference between the two agents). However, violations of injunctive norms are not *always* associated with being responsible for an outcome. And in such cases the two theories plausibly make diverging predictions.[12]

2. The Trolley Problem

In some situations, an agent does the right thing—i.e., acts fully in accord with the relevant moral norms—but nonetheless is more responsible for a negative outcome than if she had done nothing at all. One such case is the switch version of the trolley problem.[13] In the switch case, a runaway trolley is barreling down a track toward five people. An agent stands next to a switch. If the agent flips the switch, the trolley will be diverted onto a sidetrack with one person on it. Hence, the agent is faced with a decision. She can do nothing and allow five people to die, or she can flip the switch so that only the one person on the sidetrack dies. Popular sentiment holds that

[12] See Kominsky et al. (2015) and Sytsma (ms) for another type of case where the two views make diverging predictions.

[13] The trolley problem is typically understood as arising from a comparison between two cases, the switch case and the footbridge case originally given by Thomson (1985), expanding on a scenario from Foot (1978).

the agent should flip the switch. In other words, people tend to think that the moral thing for the agent to do—what the agent ought to do—is flip the switch.[14]

Despite being the morally correct action, flipping the switch has a serious, negative consequence: it results in the death of an innocent person. If not for the agent flipping the switch, the person on the sidetrack would not have died. Consequently, we expect that people will be inclined to judge that the agent is at least partially responsible for the resulting death when she flips the switch. In contrast, we expect that people will not see the agent as being comparably responsible for the death of the five when she does not flip the switch. The difference is that while the five were already in harm's way independently of the agent, the one on the sidetrack is put into harm's way by the agent's action. If the agent (through no fault or negligence on her part) had not been present, the five people would have died and the one person would have been safe.

If our judgments about how people will think about the switch version of the trolley problem are accurate, then it is an interesting test case. Both the responsibility and counterfactual views predict an asymmetry in ordinary causal attributions between the variation where the agent flips the switch and the variation where she does not. But the responsibility view predicts that the asymmetry will point in one direction, while the counterfactual view predicts that it will point in the other. Specifically, the responsibility view predicts that people will be *more* likely to say that the agent caused the outcome when she flips the switch than when she does not flip the switch,

[14] For instance, in a recent article Williamson (2016, 23) writes that "it is said to be a philosophical intuition that, in the hypothetical scenario, the subject ought to divert the trolley to save five lives at the expense of one," while Greene (2016, 175) states that it is "a matter of psychological fact" that "people responding to the standard switch case… tend to approve of hitting a switch that will redirect a trolley away from five and onto one." While many studies in the large empirical literature on judgments about trolley cases ask about whether flipping the switch is "morally permissible" (e.g., Mikhail, 2011; Liao et al., 2012) or "morally acceptable" (e.g., Greene et al., 2009), others ask what the participant would do (e.g., Petrinovich, O'Neill, and Jorgensen, 1993) or what they ought to do (e.g., Bourget and Chalmers, 2014). And in his recent book on the trolley problems, Edmonds (2014, 93) reports that "the BBC found that roughly four out of five agreed that the trolley should be diverted down the spur." We return to the question of whether people judge that one morally ought to flip the switch, as opposed to it merely being morally permissible to do so, in Section 4.

but the counterfactual view predicts that people will be *less* likely to say that the agent caused the outcome when she flips the switch than when she does not flip the switch. Let's see why.

Start with the counterfactual view. The most salient norm for the agent is the injunctive norm that she ought to flip the switch. Hence, in the case where the agent doesn't flip the switch, people should be more likely to consider the counterfactual "if the agent had flipped the switch, then the five workmen would have lived" and judge that the agent caused those deaths. By contrast, if the agent flips the switch, she is conforming with the norm and people should be less likely to consider the counterfactual "if the agent had not flipped the switch, then the one workman would have lived." So, the counterfactual view predicts that people will be more likely to say that the agent caused the bad outcome when she refrains (and thereby violates the norm) than when she flips the switch (and thereby conforms to the norm).

The responsibility view makes the opposite prediction: we predict that people will be *more* likely to say that the agent caused the outcome when she flips the switch than when she does not flip the switch. This prediction follows from the responsibility view when coupled with our guess about how people will judge the agent's normative responsibility for the outcome in each version of the scenario. On the responsibility view, we expect causal attributions to be similar to judgments of normative responsibility. And in the trolley case, we expect people will find the agent to be more responsible for the outcome when she flips the switch, since the person on the sidetrack was in no danger before the agent intervened. If our expectations about this case are accurate, then the responsibility view predicts that people will be more likely to judge that the agent caused the outcome when she flips the switch than when she does not flip the switch.

3. Testing the Trolley Case

To test the predictions laid out in the previous section, we assessed people's judgments about two variations on the trolley case—one variation in which the agent flips the switch and one variation in which she refrains from flipping the switch. The vignette for the variation in which the agent flips the switch is given below (again, full probes for all studies are available in the supplemental materials):

> A runaway trolley is headed toward five innocent people who are on the track and who will be killed unless something is done. Marcy is too far away to warn the people to get off the track, but she is standing next to a switch that she can flip to redirect the trolley onto a second track. If Marcy flips the switch, the five people on the original track will be saved. However, a bystander is standing on the second track. If Marcy flips the switch, the trolley will hit and kill the bystander.
>
> Marcy flips the switch, redirecting the trolley onto the second track. The trolley hits and kills the bystander, but not the five people.

The vignette for the other variation changes the text in the last two sentences. For each of the vignettes we asked participants about (a) whether the agent acted morally, (b) whether she was responsible for the outcome, and (c) whether she caused the outcome. We used a between-participants design. Each participant was given one of the two vignettes (flip, refrain) and then asked to assess one of the three claims—moral attribution (a), responsibility attribution (b), or causal attribution (c)—using a 7-point Likert scale anchored at 1 with "strongly disagree," at 4 with "neutral," and at 7 with "strongly agree." Responses for Study 3 were collected from 356 participants.[15] Results are shown in Figure 2.

[15] Participants were 77.0% women, with an average age of 33.7 years, and ranging in age from 16 to 79.

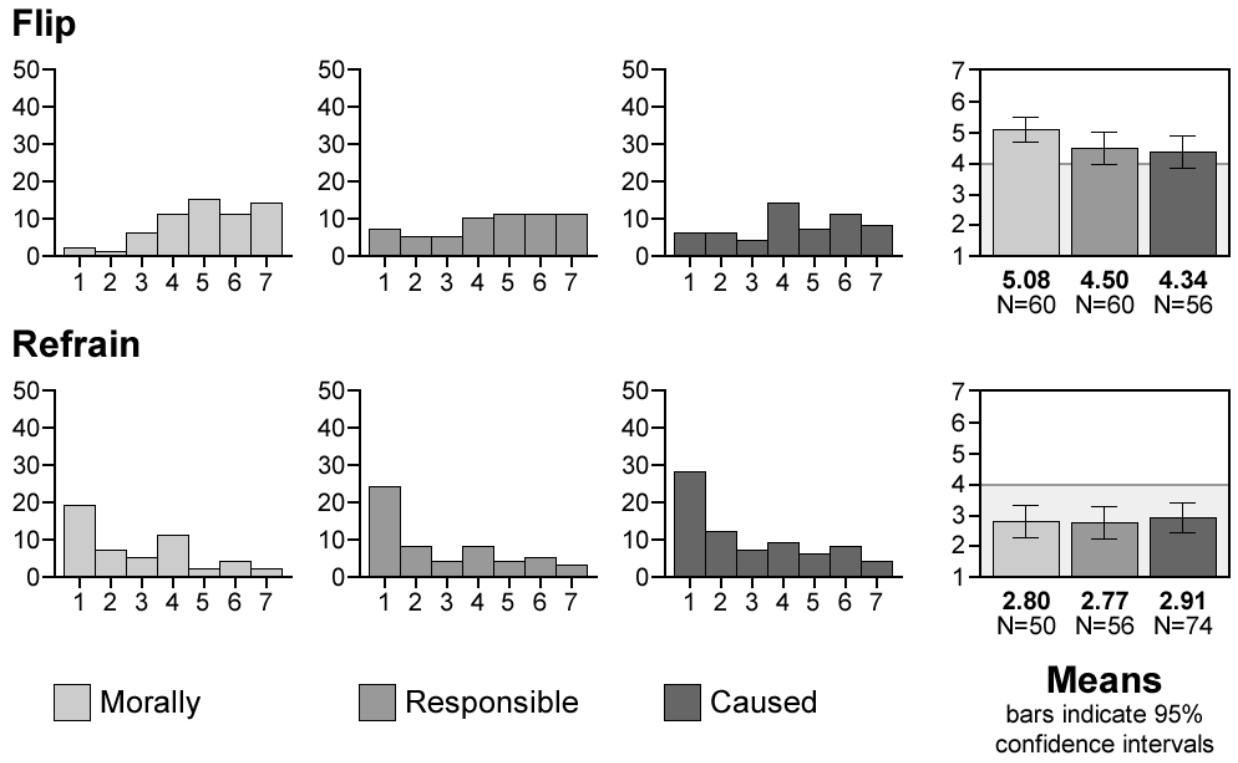

Figure 2: Results for Study 3

Causal ratings were statistically higher in the flip condition (mean=4.34 with 95% CI of [3.83, 4.85]) than in the refrain condition (mean=2.91 with 95% CI of [2.44, 3.37]).[16] And responsibility ratings (mean=4.50 with 95% CI of [3.99, 5.01]; mean=2.77 with 95% CI of [2.23, 3.30]) were statistically higher than morality ratings (mean=5.08 with 95% CI of [4.68, 5.48]; mean=2.80 with 95% CI of [2.27, 3.33]).[17] Further, causal ratings and responsibility ratings were remarkably similar.[18] This pattern of findings agrees with the predictions of the responsibility view: causal ratings are higher in the flip condition than in the refrain condition, matching the

[16] t(121.56)=4.16, p=5.9e-5, Cohen's d=0.73 (medium); W=2898, p=8.0e-5, Cliff's delta=0.40(medium). Ratings in the flip condition are not statistically different from the neutral point: t(55)=1.34, p=0.19; V=546, p=0.23. But ratings for the refrain condition are: t(73)=-4.71, p=1.2e-5; V=428, p=1.8e-5.

[17] Responsibility: t(113.19)=4.72, p=6.9e-6, Cohen's d=0.88 (large); W=2455, p=1.4e-5, Cliff's delta=0.46 (medium). Morality: t(103.19)=4.22, p=5.3e-5, Cohen's d=0.82 (large); W=2016, p=7.5e-5, Cliff's delta=0.44 (medium).

[18] Flip: t(113.84)=-0.45, p=0.65, Cohen's d=-0.083 (negligible); W=1586, p=0.60, Cliff's delta=-0.056 (negligible). Refrain: t(118.89)=0.39, p=0.70, Cohen's d=0.069 (negligible); W=2168, p=0.64, Cliff's delta=0.046 (negligible).

respective responsibility ratings. But the pattern disagrees with the predictions of the counterfactual view. As detailed above, on the counterfactual view we would expect to find an inverse relationship between morality ratings and causal ratings in each condition, resulting in causal ratings being higher in the refrain condition than in the flip condition. As such, the results for the trolley case provide prima facie evidence favoring the responsibility view over the counterfactual view.

4. Reconsidering the Counterfactual View

The responsibility view offers a simple, direct explanation of the causal ratings in Study 3. By contrast, the most natural predictions for a proponent of the counterfactual view to make do not agree with our observations. Therefore, proponents of the counterfactual view will need to produce an alternative interpretation or find auxiliary hypotheses that allow them to explain the observed asymmetry in causal ratings.

In this section, we consider four possibilities: first, one might question whether participants' morality ratings track the relevant normative evaluations; second, one might focus on the difference between acts and omissions as a non-normative factor explaining our results; third, one might argue that there is a normative difference between acting and refraining; and finally, one might note that the outcomes vary between the conditions, with one person dying in the flip condition and five people dying in the refrain condition. We'll consider each of these possibilities in turn.

4.1 Interpreting the Morality Ratings

We have assumed that when participants gave high ratings to the morality attribution—"Marcy acted morally"—they meant to indicate that Marcy did what she *ought* to have done and not

merely what she was *permitted* to do. But a proponent of the counterfactual view might argue that participants instead read the attribution in terms of moral permissibility and that judgments about permissibility do not allow us to infer that participants accept the relevant injunctive norm. While it strikes us as decidedly implausible that participants tended to read the morality attribution in this way, we do not think it would it matter if they did. The reason is that we not only found that people tended to agree with the attribution in the flip condition, but that they tended to disagree with it in the refrain condition. Let's assume for the sake of argument that participants did read the morality attribution in terms of moral permissibility. Then it would be plausible to interpret them as tending to hold that it is *impermissible* to refrain from flipping the switch. And this would seem to establish the norm at issue: the pattern of results indicates that people tend to hold that Marcy *ought* to flip the switch, whether we read the morality attribution as intended or in terms of moral permissibility.

4.2 Acts versus Omissions

The counterfactual view does not claim that injunctive norms are the *only* factors that matter for ordinary causal attributions. As such, proponents might argue that the asymmetry we found in causal ratings for the trolley case are explained by some other factor. Given that some have questioned whether omissions have the same causal status as actions, the most likely response is to note that causal ratings are higher when Marcy acts than when she refrains.[19] Perhaps our results reflect the fact that people are simply disinclined to treat omissions as causes and that this disinclination overwhelms consideration of injunctive norms in the trolley case.

[19] For theoretical discussions of causation by omission, see Schaffer (2000); Beebee (2004); Mellor (2004); Lewis (2004); McGrath (2005); Moore (2009); Bernstein (2014). For empirical work, see Livengood and Machery (2007); Wolff, Barbey, and Hausknecht (2010); Clarke et al. (2015); Henne, Pinillos and De Brigard (2017); Bello et al. (2017); Stephan et al. (2017); Khemlani et al. (2018); Willemsen (2019).

At first glance, this might not look like a promising response for advocates of the counterfactual view to put forward. After all, it seems that in the refrain condition there is both a relevant counterfactual and a relevant injunctive norm to make it salient. In fact, Henne, Pinillos, and De Brigard (2017, 273) argue that Hitchcock and Knobe's view predicts that "where omissions violate norms (and there is counterfactual dependence), these will be judged to be causes." Both of those conditions hold in the refrain condition of the trolley case. And other results in the literature (e.g., Willemsen 2019) also support the prediction that omissions in the trolley case will be judged to be causes.

Recall the variation on the e-mail case that we tested in our first study. Two agents jointly bring about a bad outcome (important e-mails deleted). One agent contributes through a norm-conforming action (logging into the computer system), while the other contributes through a norm-violating omission (failing to log out of the computer system). Despite the fact that the norm-violating agent contributed to the outcome via an *omission* while the norm-conforming agent contributed via an *action*, we found a large difference in causal ratings in this case, with participants tending to affirm that the norm-violating agent caused the problem and tending to deny that the norm-conforming agent caused the problem. Further, the results were similar in our second study when both agents contributed via omission.

The results of our first two studies demonstrate both that people are *at least sometimes* willing to attribute causation to agents who fail to do something and that the effect of norms on ordinary causal attributions also *at least sometimes* holds for omissions. In light of such findings, it seems that a compelling explanation of the effect of norms on ordinary causal attributions will need to apply to both actions and omissions. As such, for advocates of the counterfactual view to respond to our results by arguing that they simply reflect a disinclination to treat omissions as causes is not adequate. Advocates of the counterfactual view might argue that while people are

sometimes willing to treat omissions as causes, they aren't willing to do so in the trolley case. They would then need to point out some relevant difference between the trolley case and cases where people treat omissions as causes. We do not know what differences between the cases might be called on. Nor do we know what theoretical motivation might be proffered for expecting the two cases to be different. Still, the worry about the act-omission distinction seems salient enough to warrant further testing.

According to the present objection, the asymmetry in causal ratings seen in Study 3 is due to the act-omission distinction. Accepting this, the critic should then predict that the asymmetry will dissipate if we were to focus the vignettes on the shared positive occurrence in each condition—that the agent makes a decision, either the decision to flip the switch or the decision to refrain. To focus attention on the shared positive occurrence, we rewrote the vignettes from Study 3 to emphasize that the agent (Tom) was faced with a decision. As in our third study, we used a between-participants design with each participant being given one of the two vignettes. In this study, however, every participant assessed the causal attribution.[20]

We see the same asymmetry in causal ratings as we did in our third study, despite rewriting the vignette to focus on the agent's decision. Causal ratings were statistically higher when Tom decided to flip the switch (mean=5.28 with a 95% CI of [4.75, 5.80]) than when he decided not to flip the switch (mean=3.30 with a 95% CI of [2.70, 3.90]).[21] And both means were statistically different from the neutral point.[22] The observed asymmetry is in line with the prediction of the responsibility view, but it runs counter to the prediction of the counterfactual view. Further, the asymmetry did not dissipate relative to what we saw in Study 3. In fact, it deepened: in Study 3 the difference in the means for the causal ratings was 1.43 (4.34 for flip

[20] Participants were 73.0% women, with an average age of 37.3 years, and ranging in age from 17 to 97.
[21] t(105.97)=4.96, p=2.7e-6, Cohen's d=0.95 (large); W=2299, p=4.6e-6, Cliff's delta=0.50 (large).
[22] Act: t(57)=4.84, p=1.0e^{-5}; V=1033, p=8.9e-5. Refrain: t(52)=-2.34, p=0.023; V=294.5, p=0.017.

compared to 2.91 for refrain), while the difference in the present study is 1.98 (5.28 for flip compared to 3.30 for refrain).

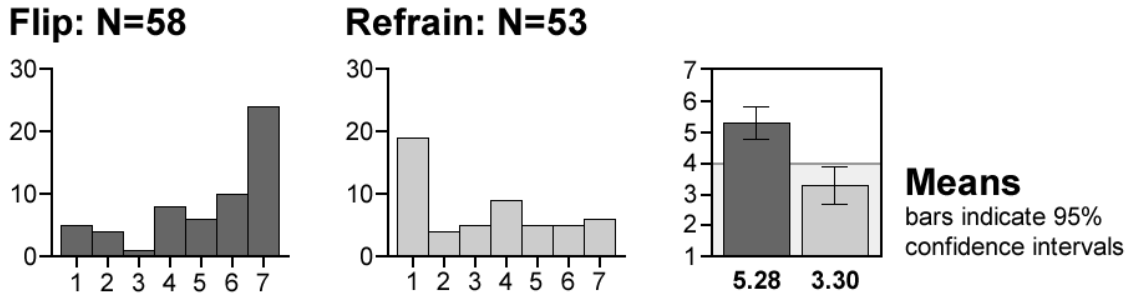

Figure 3: Results for Study 4

We did not see the asymmetry in causal ratings dissipate in our fourth study relative to our third study, as we would have expected if the act-omission objection were correct. One might argue, however, that our results show simply that we didn't sufficiently redirect attention to the shared positive occurrence. To further direct attention to the shared positive occurrence, we changed the causal statements from our fourth study to ask whether Tom's *decision* caused the outcome, while keeping everything else the same. Responses for Study 5 were collected from 110 participants.[23] Results are shown in Figure 4.

Once again, we see the asymmetry in causal ratings predicted by the responsibility view. Causal ratings were statistically higher when Tom decided to flip the switch (mean=5.34 with a 95% CI of [4.84, 5.84]) than when he decided to refrain (mean=3.62 with a 95% CI of [3.06, 4.18]).[24] And the asymmetry did not dissipate relative to what we saw in Study 3, with the difference in means for the causal ratings (1.72) being similar to what we saw previously (1.98, 1.43).

[23] Participants were 72.7% women, with an average age of 37.4 years, and ranging in age from 16 to 80.
[24] t(107.95)=4.59, p=1.2e-5, Cohen's d=0.86 (large); W=2162, p=5.6e-5, Cliff's delta=0.44 (medium)

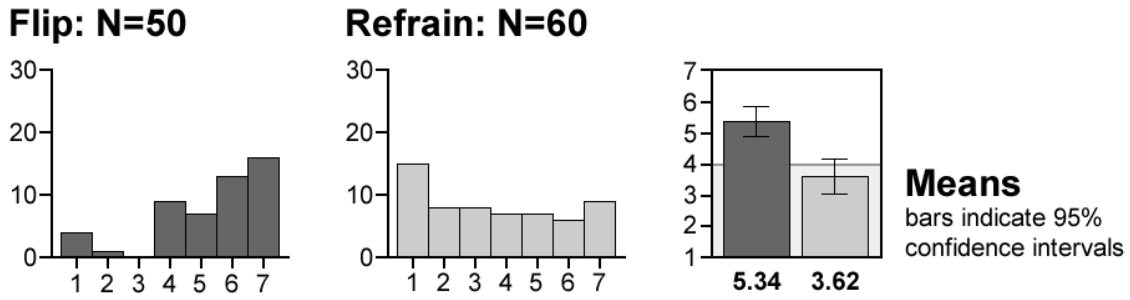

Figure 4: Results for Study 5

The results of our fifth study further bolster the case against the objection that the asymmetry in causal attributions between the flip and refrain conditions can be explained away by calling on the act-omission distinction. Focusing attention away from the distinction between acts and omissions, and toward the shared positive occurrence of the agent making a decision, did not serve to mitigate the asymmetry in causal ratings. Perhaps we've still not done enough to focus attention on the positive occurrence. To further test the objection, in our sixth study we altered the probes to remove the five people from the main track in the flip condition and to remove the person from the sidetrack in the flip condition. The result is that there is no longer a tradeoff in lives. That is, if Tom had decided to refrain in the flip condition, no one would have died; and, if Tom had decided to flip the switch in the refrain condition, no one would have died. We reasoned that if people are generally unwilling to treat omissions as causes in scenarios like this, then causal ratings in the flip condition should continue to be higher than in the refrain condition. However, that is not what we found.

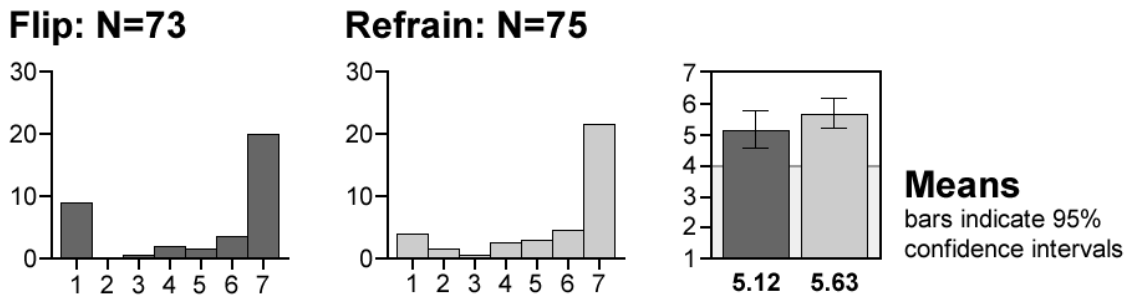

Figure 4: Results for Study 6

Responses for Study 6 were collected from 148 participants.[25] Results are shown in Figure 4. Unlike in the previous cases, the mean causal rating in the refrain condition (mean=5.63 with a 95% CI of [5.15, 6.10]) was *higher* than in the flip condition (mean=5.12 with a 95% CI of [4.53, 5.71]).[26] Further, ratings in the refrain condition were statistically *above* the neutral point.[27] This indicates that people are in fact willing to treat omissions as causes in trolley cases, just as in the e-mail cases we started with. We conclude that the act-omission distinction does not explain the asymmetry in causal ratings seen in Studies 3, 4, and 5.

4.3 Acting is Abnormal

An alternative response is to argue that our original findings for the trolley case are due to people being more likely to think that acting is abnormal, not to people being disinclined to treat omissions as causes, per se. Such a move would harken back to Hitchcock's (2007) account on which causal attributions depend on specifying default and deviant values for variables in a causal model. In that essay, Hitchcock wrote that "in the case of human actions, we tend to think of those states requiring voluntary bodily motion as deviants and those compatible with lack of

[25] Participants were 60.8% women, with an average age of 40.9 years, and ranging in age from 16 to 82.
[26] t(138.78)=-1.32, p=0.19, Cohen's d=-0.22 (small); W=2534, p=0.39, Cliff's delta=-0.074 (negligible).
[27] Flip: t(72)=3.79, p=0.00031; V=1683.5, p=0.0021. Refrain: t(74)=6.82, p=2.1e-9; V=2080.5, p=2.5e-7.

motion as defaults" (507). Hitchcock's (2007) account is readily seen as a precursor to the counterfactual view defended in Hitchcock and Knobe (2009). More recently, Knobe and colleagues (Henne et al. 2019, 158) have pointed out that "existing research suggests that people tend to be especially drawn to consider counterfactuals in which an action is replaced by an inaction rather than counterfactuals in which an inaction is replaced by an action." As such, advocates of the counterfactual view might try to explain our results by arguing that the counterfactual at issue in the refrain condition will be less likely to be considered than the counterfactual at issue in the flip condition.

Ultimately, we do not believe that this objection fairs any better than the general act-omission objection given above. While the imagined response would allow counterfactual relevance to explain our original findings for the trolley case, it would not explain the results in the follow-up studies. First, it would have trouble explaining why causal ratings weren't higher in the flip condition than in the refrain condition in Study 6. If acting is treated as abnormal enough to overwhelm the effect of the moral norm in the original study, then the counterfactual view should predict the same type of effect in the last study. But we found just the opposite. Second, if an explanation that treats positive actions as abnormal were correct, then we would expect to see the asymmetry in causal ratings for the trolley case dissipate as we focused attention on the agent's decision in Studies 4 and 5. But, again, this is not what we found.

4.4 Different Outcomes

Finally, it might be noted that the outcome in the trolley case differs depending on whether the agent flips the switch (the one person on the sidetrack dies) or whether she refrains (the five people on the main track die). And it might then be argued that determining a suitable intervention point depends on which of those outcomes needs to be prevented. The advocate of

the counterfactual view might assert that when the agent flips the switch, people will find that the best way to have prevented *the outcome of the person on the sidetrack dying* would have been for the agent to refrain from flipping the switch. In contrast, when the agent refrains, the advocate of the counterfactual view might assert that people will find that the best way to have prevented *the outcome of the five people on the main track dying* would have been to instead change something about the trolley. If these assertions are correct, then the counterfactual view would predict that causal ratings would be higher for the agent in the flip condition than in the refrain condition, just as we found.

While this is an interesting response, we do not think it is well-motivated. The main thing to note is that focusing on the separate outcomes in this way does not do justice to the situation described in the trolley case, including the tradeoff facing the agent in making her decision. But comparing the results of Studies 5 and 6 makes clear that people do recognize the dilemma. Doing so, it seems that there are two primary intervention points—an early one (changing something about the trolley) and a late one (changing the agent's decision). Changing something about the trolley, such as making it so that it was not out of control in the first place, would not simply prevent the five people on the main track from dying, however, but would remove the threat to them and with it the decision faced by the agent. As such, this intervention would undo either outcome. Changing the agent's decision, by contrast, would just switch from one outcome to the other.

To explain our pattern of findings, the advocate of the counterfactual view would need to argue that people only latch onto the early intervention point in the refrain condition in the original scenario, while instead latching onto the late intervention point in the flip condition of the original scenario and both conditions in the revised scenario used in Study 6. While this is possible, we do not see a clear theoretical motivation for this prediction. Ultimately, it seems that

the counterfactual view should either predict that causal ratings for the agent will be low in both conditions (treating the trolley as the most suitable intervention point to prevent either outcome) *or* that causal ratings will be higher when the agent refrains than when she acts (treating the agent as the most suitable intervention point to prevent the worse of the two outcomes). But we find neither pattern of results.

5. Conclusion

In this paper, we have described two potential explanations of the fact that injunctive norms matter for ordinary causal attributions and presented new empirical evidence favoring the responsibility view over the counterfactual view. In the switch version of the trolley problem, people judge that the agent ought to flip the switch and yet judge that in doing so she is more responsible for the resulting outcome than if she had refrained. And in line with the prediction given by the responsibility view, but against the predictions of the counterfactual view, people were also more likely to treat the agent as the cause of the outcome when she acted.

We observed in the introduction that metaphysicians working on causation face a dilemma. On the one hand, many philosophers think that accounts of the metaphysics of causation should be informed or constrained by ordinary intuitions, ordinary attributions, ordinary concepts, common sense, or the like. And ordinary attributions are sensitive to injunctive norms. On the other hand, injunctive norms are supposed to be irrelevant to the metaphysics of what causes what. The counterfactual view offers a way out of the dilemma by showing how ordinary causal attributions may be sensitive to injunctive norms without the ordinary, common-sense concept having any normative content. If we are right, then this way out of the dilemma is blocked. More generally, if the responsibility view is correct, then one may either appeal to ordinary intuitions and the like in order to constrain one's metaphysical account

of causation or one may develop an account according to which causation has no normative content. But one cannot do both.

References

Alicke, M. (1992). "Culpable causation." *Journal of Personality and Social Psychology*, 63: 368–378.

Alicke, M. (2000). "Culpable Control and the Psychology of Blame." *Psychological Bulletin*, 126(4): 556–574.

Alicke, M., Rose, D., and Bloom, D. (2011). "Causation, Norm Violation and Culpable Control." *Journal of Philosophy*, 108: 670-696.

Bello, P., Wasylyshyn, C., Briggs, G., & Khemlani, S. (2017). Contrasts in reasoning about omissions. In *CogSci*.

Bernstein, S. (2014). Omissions as possibilities. *Philosophical Studies*, *167*(1), 1-23.

Beebee, H. (2004). "Causing and Nothingness." In Collins, Hall, and Paul (eds.), *Causation and Counterfactuals*. MIT Press.

Bourget, D. and Chalmers, D. (2014). "What do philosophers believe?" *Philosophical Studies*, 170(3): 465-500.

Clarke, R., Shepherd, J., Stigall, J., Waller, R., and Zarpentine, C. (2015). "Causation, norms, and omissions: A study of causal judgments." *Philosophical Psychology*, *28*(2): 279-293

Collins, J., Hall, N., and Paul, L.A. (2004). "Counterfactuals and Causation: History, Problems, and Prospects." In Collins, Hall, and Paul (eds.), *Causation and Counterfactuals*. MIT Press.

Edmonds, D. (2014). *Would You Kill the Fat Man? The Trolley Problem and What Your Answer Tells Us about Right and Wrong*. Princeton: Princeton University Press.

Foot, P. (1978). *Virtues and Vices and Other Essays in Moral Philosophy*. Berkeley: University of California Press.

Greene, J. (2016). "Solving the Trolley Problem." In J. Sytsma and W. Buckwalter (eds.), *A Companion to Experimental Philosophy*. Malden: John Wiley & Sons.

Greene, J., Cushman, F., Stewart, L., Lowenberg, K., Nystrom, L., and Cohen, J. (2009). "Pushing Moral Buttons: The Interaction between Personal Force and Intention in Moral Judgment." Cognition, 111(3): 364-371.

Hall, N. (2004). Rescued from the rubbish bin. *Philosophy of Science* 71, 1107-1114.

Halpern, J. (2016). *Actual Causality*. MIT Press.

Halpern, J. and Hitchcock, C. (2015). "Graded Causation and Defaults." *The British Journal for the Philosophy of Science*, 66: 413-457.

Hart, H. and Honoré, T. (1985). *Causation in the Law*. Oxford: Oxford University Press.

Henne, P., Pinillos, Á., and De Brigard, F. (2017). "Cause by Omission and Norm: Not Watering Plants." *Australasian Journal of Philosophy*, 95(2): 270-283.

Henne, P., Niemi, L., Pinillos, Á, De Brigard, F., and Knobe, J. (2019). "A counterfactual explanation for the action effect in causal judgment." *Cognition*, 190: 157-164.

Hilton, D. and Slugoski, B. (1986). "Knowledge-Based Causal Attribution: The Abnormal Conditions Focus Model." *Psychological Review*, 93: 75-88.

Hitchcock, C. (2007). "Prevention, Preemptions, and the Principle of Sufficient Reason," *Philosophical Review*, 116(4), 495–531.

Hitchcock, C., and J. Knobe (2009). "Cause and Norm." *Journal of Philosophy*, 106: 587-612.

Khemlani, S., Wasylyshyn, C., Briggs, G., & Bello, P. (2018). Mental models and omissive causation. *Memory & cognition*, *46*(8), 1344-1359.

Knobe, J. (2006). *Folk Psychology, Folk Morality*. Dissertation.

Knobe, J. and Fraser, B. (2008). "Causal judgments and moral judgment: Two experiments." In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Volume 2: The Cognitive Science of Morality*, pp. 441–447. Cambridge: MIT Press.

Kominsky, J., Phillips, J., Gerstenberg, T., Lagnado, D., and Knobe, J. (2015). "Causal Superseding." *Cognition*, 137: 196-209.

Lewis, D. (1986). *Philosophical Papers*, Volume II. Oxford: Oxford University Press.

Lewis, D. (2004). "Causation as Influence." In Collins, Hall, and Paul (eds.), *Causation and Counterfactuals*. MIT Press.

Liao, S. M., Wiegmann, A., Alexander, J., and Vong, G. (2012). "Putting the Trolley in Order: Experimental Philosophy and the Loop Case." Philosophical Psychology, 25(5): 661-671.

Liebesman, D. (2011). "Causation and the Canberra Plan." *Pacific Philosophical Quarterly*, *92*(2): 232-242.

Livengood, J. and Machery, E. (2007). "The Folk Probably Don't Think What You Think They Think: Experiments on Causation by Absence." *Midwest Studies in Philosophy*, 31: 107–127.

Livengood, J. and Rose, D. (2016). "Experimental Philosophy and Causal Attribution." In Sytsma and Buckwalter (eds.), *A Companion to Experimental Philosophy*. Wiley Blackwell.

Livengood, J., and Sytsma, J. (forthcoming). "Actual Causation and Compositionality." *Philosophy of Science*.

Livengood, J., Sytsma, J., and Rose, D. (2017). "Following the FAD: Folk attributions and theories of actual causation." *Review of Philosophy and Psychology*, 8(2), 273-294.

McGrath, S. (2005). "Causation by Omission." *Philosophical Studies*, 123: 125-148.

Mellor, D.H. (2004). "For Facts as Causes and Effects." In Collins, Hall, and Paul (eds.), *Causation and Counterfactuals*. MIT Press.

Menzies, P. (1996). "Probabilistic Causation and the Pre-emption Problem." *Mind*, *105*(417): 85-117.

Mikhail, J. (2011). *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*. Cambridge: Cambridge University Press.

Moore, M.S. (2009). *Causation and responsibility: An essay in law, morals, and metaphysics*. Oxford University Press.

Paul, L.A. and Hall, N. (2013). *Causation: A user's guide*. Oxford University Press.

Petrinovich, L., O'Neill, P., and Jorgensen, M. (1993). "An Empirical Study of Moral Intuitions: Toward an Evolutionary Ethics." *Journal of Personality and Social Psychology*, 64: 467–478.

Reuter, K., Kirfel, L., van Riel, R., and Barlassina, L. (2014). "The good, the bad, and the timely: how temporal order and moral judgment influence causal selection." *Frontiers in Psychology*, 5: 1336.

Rose, D. (2017). "Folk Intuitions of Actual Causation: A Two-pronged Debunking Explanation." *Philosophical Studies*., 174(5): 1323-1361.

Samland, J. and M. Waldmann (2015). "Highlighting the Causal Meaning of Causal Test Questions in Contexts of Norm Violations." In D. Noelle, R. Dale, A. Warlaumont, J. Yoshimi, T. Matlock, C. Jennings, and P. Maglio (eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, pp. 2092–2097, Austin, TX: Cognitive Science Society.

Samland, J. and M. R. Waldmann (2016). "How prescriptive norms influence causal inferences." *Cognition*, 156: 164–176.

Samland, J., M. Josephs, M. Waldmann, and H. Rakoczy (2016). “The Role of Prescriptive Norms and Knowledge in Children’s and Adults’ Causal Selection.” *Journal of Experimental Psychology: General*, 145(2): 125–130.

Schaffer, J. (2000). “Causation by disconnection.” *Philosophy of Science*, 67(2): 285-300.

Stephan, S., Willemsen, P., and Gerstenberg, T. (2017). Marbles in Inaction: Counterfactual Simulation and Causation by Omission. In *CogSci*.

Sytsma, J. (ms). “The Extent of Causal Superseding.”

Sytsma, J., Bluhm, R., Willemsen, P., and Reuter, K. (2019) “Causal Attributions and Corpus Analysis.” In E. Fischer and M. Curtis (Eds.), *Methodological Advances in Experimental Philosophy*, Bloomsbury, 209-238.

Sytsma, J. and Livengood, J. (2015). *The Theory and Practice of Experimental Philosophy*. Broadview Press.

Sytsma, J., Livengood, J., and Rose, D. (2012). “Two Types of Typicality: Rethinking the Role of Statistical Typicality in Ordinary Causal Attributions.” *Studies in History and Philosophy of Science Part C*, 43: 814-820.

Thomson, J. J. (1985). “The Trolley Problem.” *The Yale Law Journal*, 94(6): 1395–1415.

Willemsen, P. (2019). *Omissions and Their Moral Relevance*. Leiden: mentis Verlag.

Williamson, T. (2016). “Philosophical Criticisms of Experimental Philosophy.” In J. Sytsma and W. Buckwalter (eds.), *A Companion to Experimental Philosophy*. Malden: John Wiley & Sons.

Wolff, P., Barbey, A.K., and Hausknecht, M. (2010). “For want of a nail: How absences cause events.” *Journal of Experimental Psychology: General*, 139(2): 191.