Intervention, Bias, Responsibility… 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 three 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 intervention view, the bias view, and the responsibility view. We then present new experimental evidence concerning a type of case not previously investigated in the literature. In the switch version of the trolley problem, people judge that the bystander 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. And, as predicted by the responsibility view, but not the intervention or bias views, people are more likely to say that the bystander caused the outcome when she flips the switch.

Keywords: actual causation, responsibility, intervention, bias, trolley problem
Intervention, Bias, Responsibility… 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. 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 growing body of research indicates that injunctive norms matter for ordinary causal attributions.

---

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; Livengood, 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 [blinded].
In case after case, we find that when an individual contributes to bringing about a negative outcome, causal ratings are higher when that individual violates an injunctive norm than when she does not. To illustrate, consider the Lauren and Jane case tested in Livengood, Sytsma, and Rose (2017):

Lauren and Jane both work for a company that uses a mainframe that can be accessed from terminals on different floors of its building. The mainframe has recently become unstable, so that if more than one person is logged in at the same time, the system crashes. Therefore, the company has instituted a temporary policy restricting the use of terminals so that two terminals are not used at the same time until the mainframe is repaired. The policy prohibits logging in to the mainframe from the terminal on any floor except the ground floor.

One day, Lauren logged in to the mainframe on the authorized terminal on the ground floor at the exact same time that Jane logged in to the mainframe on the unauthorized terminal on the second floor. Lauren and Jane were both unaware that the other was logging in. Sure enough, the system crashed.3

When participants are asked about this case, they tend to deny that Lauren caused the system to crash while affirming that Jane caused the system to crash.4 And if we remove the part of the vignette about the company policy, people tend to deny both that Lauren caused the system to crash and that Jane caused the system to crash.5

Taken at face value, these and related empirical findings suggest that according to the ordinary conception of causation, the truth conditions for causal claims do contain a moral element.6 But if so, then either we were wrong to think that injunctive norms are irrelevant to 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,

3 This case is based on one of the cases presented in Knobe (2006). Similar cases are discussed in Sytsma, Livengood, and Rose (2012), Reuter et al. (2014), and Kominsky et al. (2015).
4 Participants were asked to assess two causal claims—“Lauren caused the system to crash,” “Jane caused the system to crash”—using a 7-point scale anchored at 1 with “strongly disagree,” at 4 with “neutral,” and at 7 with “strongly agree.” The mean rating for the Lauren claim was 2.42 compared with 5.21 for the Jane claim.
5 The mean rating for both the Lauren claim and Jane claim in the non-normed version of the story was 2.70.
6 By “taken at face value,” here, we mean roughly that ordinary attributions are accurately reflecting an underlying concept that is being competently deployed.
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. According to two prominent accounts, which we call the *intervention view* and the *bias view*, injunctive norms matter to causal attributions by way of our counterfactual judgments or by way of our culpability judgments, but the concept of causation itself has no normative content.

In this paper, we present new empirical evidence in support of the responsibility view. 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. In the case we consider, 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 intervention and bias views either do not directly explain them or (worse) predict the opposite effect. Hence, metaphysicians working on causation need to choose whether to target the ordinary conception of causation or a conception of causation that is free of normative content.

Here is how we will proceed. We begin in Section 1 by describing the intervention, bias, and responsibility views and how they account for findings like those described above. In Section 2, we describe the test case we will focus on—the switch version of the trolley problem—and explain why this case is well-situated to differentiate between the three views on offer. Judgments about this case, as well as the contrasting footbridge version of the trolley problem, are tested in Section 3. Finally, in Section 4 we critically consider some alternative explanations of our findings.
1. Three Explanations

In this section, we describe the intervention, bias, and responsibility views, paying special attention to how they account for the pattern of causal attributions we see in scenarios like the Lauren and Jane case. As we said in the introduction, our aim in this paper is to offer support for the responsibility view and thereby force metaphysicians working on causation to choose between offering an account of the ordinary conception of causation and offering an account of some conception of causation that is free of normative content. That said, identifying the best explanation of the fact that injunctive norms matter for ordinary causal attributions is surely of independent philosophical interest. For example, one might want to understand the mechanism by which ordinary causal attributions are made as part of an inquiry into human nature or the nature of the mind.7

1.1 The Intervention View

In this section, we describe the articulation of the intervention view given by Hitchcock and Knobe (2009). According to Hitchcock and Knobe’s articulation of the intervention view, causal judgments serve to identify suitable intervention points, and injunctive norms 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. They don’t consider just any which way the outcome might have been prevented, however, but focus on those aspects of the situation in which something abnormal has occurred. Hitchcock and Knobe focus on overall normality

---

7 Elsewhere ([blinded]), we have argued that such projects should be understood as part of a “neutral program” in experimental philosophy.
judgments, which includes descriptive as well as injunctive norms, but in the cases we’re concerned with, the dominant norms are injunctive norms.

The basic picture, then, is that people make causal attributions through a process of counterfactual-based reasoning guided by the evaluation of norms. More carefully, according to Hitchcock and Knobe, 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. Hence, causal attributions are guided by judgments of overall normality. And the result of this process is to generate causal judgments that serve to identify socially suitable intervention points. As such, norms play a significant role in causal cognition, but the ordinary concept of causation is not itself treated as a normative concept. Rather, injunctive norms are part of the background against which many causal judgments are made: injunctive norms come into play when people attempt to identify suitable intervention points in normatively-laden situations.

Let’s illustrate this explanation by applying it to the Lauren and Jane case. Recall that in the variation in which no policy is specified, people don’t identify either Lauren or Jane as causing the system to crash. The intervention view can interpret these judgments as indicating that neither person is seen as being a suitable intervention point. And this seems plausible. After all, neither Lauren nor Jane did anything unusual, so it seems reasonable to think that the best way to prevent the system from crashing in the future is not to specifically target either of them,

---

8 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 Sludoski (1986).
but to either change the mainframe or implement a policy restricting access to it. In contrast, when there is a policy and Jane violates the injunctive norm it establishes, her doing so marks a clear point for intervention: an obvious way to prevent the system from crashing in the future is to make sure that people follow the policy. And, as we saw above, people tend to say that Jane caused the crash in this scenario.

1.2 The Bias View

The second type of explanation we’ll consider is the bias view. We’ll focus on the version of this view developed by Alicke, Rose, and Bloom (2011) in responding to Hitchcock and Knobe. Unlike the intervention view, the bias view treats the sensitivity of ordinary causal attributions to injunctive norms as reflecting systematic performance error. The basic idea is that in the process of arriving at causal judgments, prior evaluations play an important role. Specifically, the bias view holds that ordinary causal judgments are often implicitly shaped to rationalize or validate our desires to blame or praise. We’ll refer to such desires jointly as personal evaluations. Thus, Alicke, Rose, and Bloom (2011, 670) write that “when people are asked to identify, for example, the primary cause of an event, they accord privileged status to actions that arouse positive or negative evaluations,” which has the result that “causal attributions reflect a desire to praise or denigrate those whose actions we applaud or deride.”

According to the bias view, the effect seen for the Lauren and Jane case is not directly driven by the injunctive norm, but by the corresponding personal evaluations—by the resulting desire to blame Jane and praise Lauren—with these evaluations then biasing people’s causal judgments. The judgment that Jane caused the system to crash, for instance, reflects that people desire to denigrate her for doing something she shouldn’t have—for knowingly violating a policy put in place to prevent a harmful outcome from occurring. And this desire to denigrate her then
shapes their causal judgments: since people shouldn’t be blamed for things they didn’t cause, Jane’s role in bringing about the outcome is exaggerated so as to validate the prior blame judgment.

1.3 The Responsibility View

The intervention and bias views are similar in ultimately treating the ordinary concept of causation as a descriptive concept. While injunctive norms often impact the application of the concept, either because they are relevant to identifying appropriate counterfactuals to evaluate or because they bias our judgments, the concept itself is not seen as normative. By contrast, on our responsibility view, the ordinary concept of causation has built-in normative content.9 While space prevents us from providing a detailed account of this view here (but see [blinded] for further details), a brief explanation is called for.

On our view, when “caused” is used in ordinary English it generally serves 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, whether for good or for ill. Putting this in terms of concepts, rather than the corresponding terms, we hold that the dominant ordinary concept of causation is relevantly akin to

9 Perhaps the ordinary concept of causation is a “thick” ethical concept, like the concept of cruelty (see Putnam, 2002). Of course, the practical consequences of having, as opposed to lacking, normative content depend on what the normative content is and on how that content is structured. If the concept of causation merely has normative content sufficient to make causal attributions sensitive to injunctive norms, then a norm-laden account will be difficult if not impossible to distinguish from the intervention and bias views we are considering. However, on the responsibility view, the ordinary concept of causation doesn’t merely have normative content that makes causal attributions sensitive to injunctive norms, it has normative content that leads to causal attributions generally being similar to responsibility attributions.
the ordinary concept of responsibility.\textsuperscript{10} While the ordinary concept of responsibility itself stands in need of empirical elucidation, we will treat it as a primitive for purposes of this paper, taking responsibility attributions to express normative evaluations. The responsibility view then predicts that causal attributions will generally be similar to responsibility attributions.

The responsibility view readily explains the responses seen for the Lauren and Jane case. When the agents do not violate an injunctive norm by logging in—either because there was no policy prohibiting it or because they acted in accord with the policy—people tend to hold that neither was responsible for the outcome. And, likewise, they tend to judge that neither caused the outcome. But when Jane violates the policy by logging in, people tend to hold that she is responsible for the outcome. And they also tend to judge that Jane caused the outcome.\textsuperscript{11}

2. The Trolley Problem

Despite their differences, the intervention, bias, and responsibility views arguably make the same predictions about a wide range of cases, including prominent cases from the literature like the Lauren and Jane case. The reason is that in cases like this violations of injunctive norms tend to correspond with personal evaluations, and each tends to correspond with responsibility judgments. However, violations of injunctive norms and personal evaluations are not always associated with being responsible for an outcome. In this section we describe one such case.

\textsuperscript{10} Alternatively, it might be that there is a single underlying concept, which could be thought of as causal-responsibility, that gets employed in using both “cause” and “responsible.” The philosophical concepts could then be seen as distinguishing two aspects of this more general concept.

\textsuperscript{11} Responsibility attributions for the Lauren and Jane cases were similar to the causal attributions reported above. When Jane violates the policy, the mean responsibility rating was 5.44 (compared to 5.21), and the mean responsibility rating for Lauren was 1.75 (compared to 2.42). When there is no policy, the mean responsibility rating was 2.32 for Lauren and 2.51 for Jane (compared to 2.70 for each). See [blinded] for details.
2.1 Switch Case

In some situations, an agent doing the right thing—acting in accord with the relevant moral norm and deserving praise for her action—can leave her 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. In this case, a runaway trolley is barreling down a track toward five people. A bystander is standing next to a switch. If the bystander flips the switch, the trolley will be diverted onto a sidetrack with one person on it. The bystander is faced with a decision: do nothing and allow the five to die, or flip the switch and save the five at the expense of the one. Popular sentiment holds that the bystander should flip the switch. In other words, it seems that people tend to see this case as involving a clear injunctive norm—the moral thing for the bystander to do is to flip the switch.

There is a serious consequence to flipping the switch, however, as it results in the death of an innocent person who was in no danger prior to the bystander’s action. If not for the bystander flipping the switch, the person on the sidetrack would not have died. Because of this, it seems to us (the authors) that if the bystander flips the switch, she is at least in part responsible for the death that results. In contrast, it does not seem to us that the bystander would be comparably responsible for the death of the five if she were to do nothing. The difference is that while the five were already in harm’s way independently of the bystander, the one is put into

---

12 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).
13 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, below.
14 This is not to say that she would bear no responsibility for the death of the five if she were to do nothing, but that she should bear more responsibility for the death of the one if she flipped the switch than she would be for the death of the five if she did not flip the switch.
harm’s way by her flipping the switch. After all, if the bystander (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 the switch version of the trolley problem are representative, then it is an interesting test case for the intervention, bias, and responsibility views. While each view plausibly predicts that we’ll find an asymmetry in ordinary causal attributions between the case where the bystander flips the switch and the case where she does not, the responsibility view predicts that the asymmetry will point in one direction, and the intervention and bias views predict that it will point in the opposite direction.

It seems that the most salient norm for the bystander in the switch case is the injunctive norm that she ought to flip the switch. Focusing on this norm, the simplest and most straightforward application of the intervention view generates the prediction that people will be more likely to say that the bystander caused the outcome when she violates the norm (doesn’t flip the switch) than when she acts in accordance with the norm (flips the switch). Similarly, with regard to the bias view, we would expect people to feel more disapprobation for the bystander’s action when she violates the norm than when she acts in accordance with it and more praise for her action when she acts in accordance with the norm than when she violates it. As such, the bias view would also seem to predict that people will be more likely to say that the bystander caused the outcome when she does not flip the switch than when she does. The responsibility view, by contrast, makes the opposite prediction. We expect that causal judgments will be similar to responsibility judgments; thus, if we are correct that people will be more likely to judge that the bystander is responsible for the outcome when she flips the switch than when she does not flip the switch, then we would expect people to also be more likely to judge that the bystander caused the outcome when she flips the switch than when she does not flip the switch.
2.2 Footbridge Case

To help place these competing predictions in context, it is worth considering an alternative trolley case where the three views make the same prediction. In the footbridge case, instead of standing next to a switch, the bystander is now on a footbridge going over the track. Standing next to her is a very large man. The only way that the bystander can stop the trolley before it hits the five people is to push the large man off the footbridge so that he lands on the track in front of the trolley, killing him in the process. Most people’s judgments now run counter to what we saw for the switch case: people tend to say that the bystander should not act. Thus, there again seems to be a clear injunctive norm, but this time it pushes in the opposite direction.

With the shift in injunctive norm, we also get a shift in the predictions made by the intervention and bias views. Since the bystander now violates the norm when she acts, the intervention view would seem to predict that people will be more likely to say that the bystander caused the outcome when she pushes the man than when she does not push him. And since we would expect people to feel disapprobation for the bystander’s action when she pushes the man, the bias view should make the same prediction. We do not get a corresponding shift for the responsibility view, however. With regard to responsibility judgments, it seems that the same basic reasoning holds for the footbridge case as for the switch case: when the bystander pushes the man she is more responsible for his death than she would be for the death of the five if she were to do nothing. Thus, the responsibility view also predicts that people will be more likely to say that the bystander caused the outcome when she pushes the man than when she refrains.

---

15 This is not to say that there is no difference between the cases, but simply to note that we again expect responsibility judgments to be higher when the bystander acts than when she refrains. That said, we expect that people will tend to believe that the bystander is more responsible for the death when she pushes the large man in the footbridge case than when she flips the switch in the switch case.
3. Testing the Trolley Problem

To test the predictions laid out in the previous section, we want to assess people’s judgments about two variations on each of the two trolley cases (switch, footbridge)—one variation in which the bystander acts and one variation in which she refrains. To do this we wrote four vignettes. The vignette for the acts condition for the switch case is given below (all of the probes for the studies in this paper 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.

For each of the vignettes we wanted to assess people’s judgments about (a) whether the agent acted morally, (b) whether she was responsible for the outcome, and (c) whether she caused the outcome. This was done using a between-participants design. Each participant was given one of the four vignettes, then asked to assess one of three claims—a morality claim, a responsibility claim, or a causal claim—using a 7-point Likert scale anchored at 1 with “strongly disagree,” at 4 with “neutral,” and at 7 with “strongly agree.” The three claims for the acts conditions for the switch case are:

- Marcy acted morally.
- Marcy is responsible for the death of the bystander.
- Marcy caused the death of the bystander.

Responses for Study 1 were collected from 702 participants; the results are shown in Figure 1.16, 17

---

16 In all studies, responses were collected online through [blinded]; 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, or have taken graduate-level courses in philosophy.

17 Participants were 76.6% women, with an average age of 33.8 years, and ranging in age from 16 to 79.
Figure 1: Results for Study 1
3.1 Interpreting the Morality Ratings

On the intervention and bias views we would expect to find an inverse relationship between morality ratings and causal ratings across the acts condition and the refrains condition for each of the two cases. When morality ratings are lower, we would expect causal ratings should be higher, and vice versa. And this is exactly what we found for the footbridge case: morality ratings were lower and causal ratings were higher when Marcy acted than when she refrained. This is not what we found for the switch case, however. In fact, we found the opposite: both morality ratings and causal ratings were higher when Marcy acted than when she refrained. Given this, one obvious response for a proponent of the intervention or bias view to make is to question whether the morality ratings track the relevant evaluations. It might be argued, for instance, that asking whether Marcy acted morally is ambiguous between asking whether she did what she morally ought to do or merely something that she was morally permitted to do, with the latter but not the former establishing a relevant injunctive norm.

The primary thing to note, here, is that we didn’t simply find that people affirmed that Marcy acted morally in the acts condition of the switch case, but that they also denied that she acted morally in the refrains condition. And this asymmetry is not readily interpreted in terms of mere permissibility. Suppose that people read the statements in terms of moral permissibility. It would then seem that they thought that Marcy was not morally permitted to refrain from pulling the switch. But this would then seem to imply that she morally ought to do or merely something that she was morally permitted to do, with the latter but not the former establishing a relevant injunctive norm at issue.

To further test that the morality ratings correspond with an injunctive norm and personal evaluations for the switch case, in our second study we gave participants one of the two switch vignettes used in our previous study, but changed the questions. This time we asked each participant to assess three claims using the same scale as before:
Marcy did the right thing.
Marcy is praiseworthy.
Marcy is blameworthy.

We then asked participants the following question: “Overall, how would you evaluate Marcy’s behavior?” They answered on a 7-point Likert scale anchored at 1 with “very bad,” at 4 with “neither good nor bad,” and at 7 with “very good.” Responses for Study 2 were collected from 104 participants; the results are shown in Figure 2.\footnote{Participants were 69.2\% women, with an average age of 34.6 years, and ranging in age from 16 to 70.}

In line with our expectations, we found that participants tended to affirm that Marcy did the right thing when she acted (mean=5.11 with a 95\% CI of [4.65, 5.56]), while also affirming that she was praiseworthy (mean=4.55 with a 95\% CI of [4.09, 5.01]) and denying that she was blameworthy (mean=2.62 with a 95\% CI of [2.25, 2.98]). Further, they tended to give positive evaluations of her behavior (mean=5.00 with a 95\% CI of [4.65, 5.35]).\footnote{Each comparison was significantly different from the neutral point. Right: }\footnote{Each comparison was significantly different from the neutral point except the blame rating, which was borderline significant: Right: t=-5.6267, df=56, p=6.103e^{-7}. Praise: t=-5.8218, df=56, p=2.958e^{-7}. Blame: t=-1.6281, df=56, p=0.1091. Behavior: t=-3.3091, df=56, p=0.001641.} In contrast, participants tended to deny that Marcy did the right thing when she refrained (mean=2.70 with a 95\% CI of [2.24, 3.16]). Participants also tended to deny that she was praiseworthy (mean=2.77 with a 95\% CI of [2.35, 3.19]). However, despite saying that Marcy did not do the right thing, participants were reluctant to blame her (mean=3.53 with a 95\% CI of [2.94, 4.11]), although responses tended to be higher than they were for the act condition, as we expected. Overall assessments of her behavior were negative on average (mean=3.40 with a 95\% CI of [3.04, 3.76]), but by far the most frequent answer was a ‘4,’ which was anchored at “neither good nor bad.” Overall, these findings offer further evidence that people find there to be an injunctive norm to act in the switch case and that this norm corresponds with their personal evaluations.
Figure 2: Results for Study 2
4. Alternative Predictions

As we saw in Study 1, there is an asymmetry in causal ratings between the acts condition and the refrains condition for both the switch case and the footbridge case: for each case, causal ratings are higher when Marcy acts than when she refrains. Moreover, we have seen that responsibility ratings exhibit the same basic pattern, with Marcy being judged to be more responsible when she acts than when she refrains. Hence, the responsibility view offers a simple, direct explanation of the causal ratings in both cases. The same cannot be said for the intervention and bias views, however. On the most natural interpretations of these views, the causal ratings would be explained *if* the morality ratings were always lower when Marcy acts than when she refrains, since on those views, being out of norm explains higher causal ratings. But morality ratings are *not* always lower when Marcy acts. In the switch case, participants report lower morality ratings when she refrains.

The most natural prediction for a proponent of either the intervention or the bias view to make about the switch version of the trolley problem do not agree with our observations. Therefore, in order to defend one of these views, the proponent needs to produce an alternative interpretation or find auxiliary hypotheses that would allow the asymmetry in causal ratings that we observed to be explained. It is clear how the responsibility view can accommodate our finding, since there is a relevant asymmetry in responsibility judgments about the cases, but what explanatory asymmetry could be called on by the proponent of the intervention or bias view? We see three possibilities: first, one might focus on the difference between acts and omissions as a non-normative factor explaining our results; second, one might argue that there is a normative difference between acting and refraining; finally, one might note that the outcomes vary between the conditions, with one person dying in the acts conditions and five people dying in the refrains conditions. We’ll consider each of these possibilities in turn.
4.1 Acts versus Omissions

Neither the intervention view nor the bias view claims that injunctive norms or personal evaluations are the only factors that matter for ordinary causal attributions. As such, it might be argued that the asymmetries we found in causal ratings for the trolley cases are explained by another factor all together. Given that some have questioned whether omissions have the same causal status as actions, the most plausible response is to note that in each case causal ratings are higher when Marcy acts than when she refrains.21 Thus, a proponent of the intervention or bias view might assert that our results reflect the fact that people are more inclined to treat actions as causes than they are to treat omissions as causes and that their preference for acts over omissions largely overwhelms consideration of injunctive norms or the biasing effect of personal evaluations in these cases.

If ordinary people had a strong inclination to treat actions but not omissions as causes, then our results would be consistent with a secondary effect that follows moral ratings and aligns with the intervention and bias views. The idea here is that what we should focus on is not comparing the acts condition of the switch case with the refrains condition of the switch case (and similarly for the footbridge case) but comparing the acts condition of the switch case with the acts condition of the footbridge case and the refrains condition of the switch case with the refrains condition of the footbridge case. Doing so, we find that for each of these two comparisons causal ratings were higher when morality ratings were lower, as we would expect on either the intervention view or the bias view. Thus, looking at the acts conditions, the mean morality rating for the switch case is much higher than for the footbridge case (5.08 versus 2.37) and this relationship is inverted for the causal ratings, with the mean causal rating for the switch

case being lower than for the footbridge case (4.34 versus 5.29). Similarly, for the refrains conditions: the mean morality rating is much lower for the switch case than for the footbridge case (2.80 versus 5.24) and this relationship is inverted for the causal ratings, with the mean causal rating for the switch case being higher than for the footbridge case (2.91 versus 1.93).

The core idea for this objection is that the asymmetry between the causal ratings in the acts condition and the refrains condition for the switch case can be explained away by focusing on the distinction between acts and omissions. If this is correct, then we would expect to see the asymmetry dissipate if we 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 not flip the switch. To focus attention on the shared positive occurrence, we rewrote the four vignettes from Study 1 to emphasize that the agent (Tom) was faced with a decision. As in our first study, we used a between-participants design with each participant being given one of the four vignettes. In this study, however, each participant assessed a causal claim using the same 7-point scale as in our previous studies. We then tested normative judgments on a second page. Expanding on the findings from Study 2, this time we asked participants to assess the claim that “Tom made the correct decision.” Responses for Study 3 were collected from 218 participants; the results are shown in Figure 3.

---

23 Again, both comparisons are statistically significant (two-tailed). Moral: t=-6.6076, df=105.321, p=1.642e-9. Cause: t=2.9357, df=121.663, p=0.00398.
24 We also asked participants to assess that claim that “Tom’s decision was morally permissible” and to assess a relevant counterfactual (e.g., “if Tom had decided to do nothing, the five people on the original track would have died”). As the results to these questions revealed nothing surprising, we will exclude them from the subsequent analysis.
25 Participants were 70.6% women, with an average age of 36.1 years, and ranging in age from 16 to 97.
The key thing to note is that we see the same asymmetry in causal ratings for the switch case as we did in our first study, despite rewriting the vignette to focus on the agent’s decision. Causal ratings were 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]).26 And both means were significantly different from the neutral point.27 The observed asymmetry is in line with the prediction given for the responsibility view in Section 2, but it runs counter to the predictions given for the intervention and bias views. Further, against what we would expect based on the present objection, the asymmetry did not dissipate in this study relative to what we saw in Study 1. In fact, it deepened: in the previous study the difference in

---

26 The difference was statistically significant: t= 4.9597, df=105.965, p=2.701e-6. A 95% confidence interval for the difference in means is [1.18, 2.76]. Cohen’s d is 0.95, which is a large effect.
27 Acts: t=4.8415, df=57, p=1.023e−5. Refrains: t=−2.3406, df=52, p=0.02312.
the means for the causal ratings for the switch case was 1.43 (4.34 for acts compared to 2.91 for refrains), while the difference for the switch case in the present study is 1.98 (5.28 for acts compared to 3.30 for refrains). Turning instead to comparing the results across the switch and footbridge cases, while the mean causal ratings are again lower for the acts condition of the switch case than for the acts condition of the footbridge case (5.28 versus 5.50) and higher for the refrains condition of the switch case than for the refrains condition of the footbridge case (3.30 versus 2.70), the differences are reduced from what we saw in Study 1.28

We did not see the asymmetry in causal ratings dissipate in our third study, as we would have expected if the objection were correct. It might be argued, however, that this simply reflects that we didn’t sufficiently redirect attention to the shared positive occurrence. To further direct attention in this direction, in our fourth study we changed the causal statements from our third study to ask whether Tom’s decision caused the outcome, while keeping everything else the same. Responses for Study 4 were collected from 222 participants; the results are shown in Figure 4.29

Once again we see the asymmetry in causal ratings for the switch case predicted by the responsibility view. Causal ratings were higher when Tom decided to flip the switch (mean=5.34 with a 95% CI of [4.84, 5.84]) than when he decided not to flip the switch (mean=3.62 with a 95% CI of [3.06, 4.18]).30 And once again the asymmetry did not dissipate relative to what we saw in the previous studies, with the difference in means for the causal ratings (2.36) being larger than what we saw in both Study 3 (1.99) and Study 1 (1.43). Turning to a comparison of the results across the two cases, we now find that the mean causal ratings are higher for the acts

---

28 Neither comparison is statistically significant. Acts: t=-0.607, df=109.965, p=0.5451, and a 95% confidence interval is given by [-0.96, 0.51]. Refrains: t=1.4584, df=103.849, p=0.1477, and a 95% confidence interval is given by [-0.22, 1.42].

29 Participants were 73.0% women, with an average age of 36.8 years, and ranging in age from 16 to 80.

30 The difference was statistically significant: t=4.5875, df=107.952, p=1.215e-5. A 95% confidence interval for the difference in means is [0.98, 2.47]. Cohen’s d is 0.86, which is a large effect.
condition for the switch case than for the acts condition for the footbridge case (5.34 versus 4.95), although they remain higher for the refrains condition for the switch case than for the refrains condition for the footbridge case (3.62 versus 3.11). \(^{31}\)

![Graphs showing results for Study 4](image)

**Figure 4:** Results for Study 4

The results of our fourth study further bolster the case against the objection that the asymmetry in causal attributions we found between the acts and refrains 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 bystander making a decision, did not serve to mitigate the asymmetry in causal ratings for the

\(^{31}\) Neither comparison is statistically significant. Acts: t=0.99, df=104.344, p=0.3245, and a 95% confidence interval is given by [-0.39, 1.18]. Refrains: t=1.2662, df=111.488, p=0.2081, and a 95% confidence interval is given by [-0.29, 1.30].
switch case. Further, while a proponent of the intervention or bias view could claim that the results of our first study showed a clear secondary effect that could be attributed to norm violation or personal evaluations, this is not the case for our subsequent studies. Most notably, the relationship between causal ratings and normativity ratings run in the opposite direction in the acts conditions in our fourth study—causal ratings are lower for the footbridge case than for the switch case, and the same holds for normativity ratings. Overall, across these studies there is little reason to think that the injunctive norm or corresponding personal evaluations have any effect on causal ratings.

Recall that we introduced the present objection by noting that neither the intervention view nor the bias view claims that injunctive norms or personal evaluations are the only factors that matter for ordinary causal attributions. We then suggested that a proponent of one of these views might argue that injunctive norms or personal evaluations have a secondary effect on the causal attributions for the trolley cases that is largely overwhelmed by the difference between acts and omissions. But our subsequent studies suggest that there is no such secondary effect. At this point, the proponent might stick to her guns and argue that her view has nothing directly to say about this case and makes no predictions about it one way or the other. But to defend the intervention or bias view in such a way would be pyrrhic, for without a compelling account of the conditions under which we should expect salient injunctive norms to have or not to have an effect on ordinary causal attributions, remaining silent on the trolley cases deprives the intervention and bias views of all explanatory power.

4.2 Acting is Abnormal

An alternative response is to argue that the effects we’ve seen are due not to people being more willing to treat actions as causes than omissions, per se, but that people are more likely to think
that acting is abnormal or that disapprobation is more likely to attach to our acts than to our
omissions. Making this move, the proponent of the intervention or bias view cannot be accused
of trying to explain away the effects in terms of a non-normative difference, and she might be
thought to have further resources at her disposal for explaining the pattern of responses seen
across our studies.

Ultimately, we do not believe that this objection fairs any better than the previous one.
The basic reason is that it is unclear why we should think that acting in the switch case would be
thought to be less normal or to incur greater disapprobation than refraining. The case against the
latter is especially clear, since we’ve seen that people not only think it is more moral to flip the
switch than to refrain (Study 1), but that this is both the right thing to do (Study 2) and the
correct thing to do (Study 3), that doing so is more praiseworthy and less blameworthy than
refraining (Study 2), and that participants’ overall evaluations of the behavior favor acting (Study
2). Given this, it would simply be incredible to suppose that people nonetheless feel greater
disapprobation for the bystander when she acts than when she refrains.

With regard to the intervention view, it might be argued that we should take a cue from
some earlier work of Hitchcock’s (2007), expanding the discussion of norms in Hitchcock and
Knobe (2009) to include that it is abnormal to act. Hitchcock (2007) gives an account of actual
causation that depends on specifying default and deviant values for variables in a causal model.
This is readily seen as a precursor to the abnormality judgments called on by the intervention
view. And, in that article he 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 motion as
defaults” (507). Applying this to the intervention view, it could be argued that despite the moral
norm involved in the switch case, it is still more normal for the agent to refrain than to act. But,
again, if this were correct then we would expect to see the asymmetry in causal ratings for the
switch case dissipate as we focused attention on the agent’s decision, as we did in the studies in the previous sub-section. This is not what we found, however. Further, at the heart of the intervention view is the idea that the goal is to identify suitable intervention points. The evaluations elicited in our studies, however, would seem to make clear that participants feel that the outcome in which the person on the sidetrack dies is the preferable of the two outcomes. Thus, it seems that the only suitable intervention with regard to the bystander is to have her flip the switch when she refrains.

4.3 Different Outcomes

Finally, it might be noted that the outcome in the switch case differs depending on whether the bystander acts (the one person on the sidetrack dies) or whether she refrains (the five people on the main track die). And it might be urged that determining a suitable intervention point depends on which of those outcomes needs to be prevented or that disapprobation will depend on the outcome at issue. When the bystander acts, for example, it might be thought that the best way to have prevented the outcome of the person on the sidetrack dying would have been for the bystander to refrain from flipping the switch. And since the bystander did not do this, it might be thought that people will feel disapprobation for the bystander’s action with regard to the death of the man on the sidetrack. In contrast, when the bystander refrains, it might be thought 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. And, thus, it might be thought that people will not tend to feel disapprobation for the bystander’s action with regard to the death of the five people on the main track.

While this is an interesting objection, we do not think it holds up. With regard to the intervention view, the main thing to note is that changing something about the trolley would also
prevent the person on the sidetrack from dying in the eventuality that the bystander still flips the switch, and the bystander flipping the switch would also prevent the five people on the main track from dying. As such, it is unclear that focusing on the difference in outcomes enables the intervention view to explain our findings. Again, it seems that the intervention view should either predict that causal ratings for the bystander 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 bystander refrains than when she acts (treating the bystander as the most suitable intervention point to prevent the worse of the two outcomes). With regard to the bias view we again have the same basic response as above: it seems that people are more likely to feel disapprobation for the bystander’s action when she flips the switch than when she does not (and the reverse with regard to praise), so it does not seem that their personal evaluations selectively focus on the trolley when the outcome is the death of the five rather than the death of the one.

5. Conclusion

In this paper, we have described three potential explanations of the fact that injunctive norms matter for ordinary causal attributions and presented new empirical evidence favoring the responsibility view over the intervention and bias views. In the switch version of the trolley problem, people judge that the bystander 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 most plausible predictions for the intervention and bias views, people were also more likely to treat the bystander as the cause of the outcome when she acted.

If the responsibility view is the correct account of ordinary causal attributions, then 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, etc.). On the other hand, injunctive norms are supposed to be irrelevant to what causes what. Hence, one may either appeal to ordinary intuitions and the like in order to constrain one’s account of causation or one may develop an account according to which causation has no normative content. But one cannot do both.

References


