The Character of Causation: Investigating the Impact of Character, Knowledge, and Desire on Causal Attributions

Justin Sytsma[1]

Abstract: There is a growing consensus that norms matter for ordinary causal attributions. This has important implications for philosophical debates over actual causation. Many hold that theories of actual causation should coincide with ordinary causal attributions, yet those attributions often diverge from the theories when norms are involved. There remains substantive debate about why norms matter for causal attributions, however. In this paper, I consider two competing explanations—Alicke's *bias view*, which holds that the impact of norms reflects systematic error (suggesting that ordinary causal attributions should be ignored in the philosophical debates), and our *responsibility view*, which holds that the impact of norms reflects the appropriate application of the ordinary concept of causation (suggesting that philosophical accounts are not analyzing the ordinary concept). I investigate one key difference between these views: the bias view, but not the responsibility view, predicts that "peripheral features" of the agents in causal scenarios—features that are irrelevant to appropriately assessing responsibility for an outcome, such as general character—will also impact ordinary causal attributions. These competing predictions are tested for two different types of scenarios. I find that information about an agent's character does not impact causal attributions on its own. Rather, when character shows an effect it works through inferences to relevant features of the agent. In one scenario this involves inferences to the agent's knowledge of the likely result of her action and her desire to bring about that result, with information about knowledge and desire each showing an independent effect on causal attributions.

An expanding body of evidence indicates that norms, especially injunctive norms, matter for ordinary causal attributions.[2] This has notable implications for philosophical debates concerning actual causation. Many participants in this debate are committed to what Livengood et al. (2017) term the *folk attribution desideratum* (FAD): they hold that an important measure of the

[1] I want to thank Jonathan Livengood for extremely helpful feedback on previous drafts of this paper. I would also like to thank the audience at the University of Waikato.

[2] See Hilton and Slugoski (1986), Alicke (1992), Knobe and Fraser (2008), Hitchcock and Knobe (2009), Sytsma et al. (2012), Reuter et al. (2014), Kominsky et al. (2015), Livengood et al. (2017), Icard et al. (2017), Rose (2017), Kominsky and Phillips (2019), and Livengood and Sytsma (forthcoming), among others. By "ordinary causal attributions" I specifically mean the use of language like "X caused Y." Some researchers arguably go further than this, holding that norms matter not just for ordinary causal attributions, but for causal cognition more generally (Danks et al., 2014). I will restrict myself to the narrower claim, however. In addition, I will focus on just injunctive norms, which I take to include both prescriptive norms (what should be done) and proscriptive norms (what should not be done). In the existing literature, the phrase "prescriptive norm" is often used to refer to both prescriptive and proscriptive norms.

acceptability of theories of actual causation is how well they accord with ordinary causal attributions about specific cases. The empirical literature shows, however, that ordinary causal attributions often diverge from the verdicts of philosophical theories for scenarios involving norm violations (e.g., Livengood et al. 2017, Rose 2017, Livengood and Sytsma forthcoming), raising a problem for theorists embracing the FAD.

One prominent explanation of the role of norms in ordinary causal attributions—the *bias view* put forward by Alicke and colleagues—holds that their impact owes to a systematic performance error, with the desire to blame biasing otherwise descriptive causal judgments.[3] Rose (2017) has recently drawn out the implications of this view for the philosophical debates, putting forward a debunking explanation of ordinary causal attributions. He concludes that "in the dispute over actual causation, folk intuitions deserve to be rejected" (1323). With regard to judgments about agents, this follows from the view that the problematic intuitions are the result of bias.[4] In contrast, our competing *responsibility view* holds that the impact of norms issues not from bias, but from the correct application of the ordinary concept of causation at play in causal attributions.[5] If this is correct, then insofar as the dispute over actual causation concerns the ordinary concept, folk intuitions should not be ignored.[6]

[3] See Alicke (1992, 2000), Alicke and Rose (2010), Alicke et al. (2011), Rose (2017).

[4] While Rose offers a two-pronged debunking argument, calling on the bias view with regard to judgments about agents and a separate process based on primitive teleological considerations for judgments about non-agents, I'll focus on just judgments about agents here. In doing so, it is worth noting that one prong of Rose's argument could hold without the other. For instance, it could be that people make a mistake in applying the ordinary concept of causation to non-agents, and yet apply it appropriately in making judgments about agents. This would not warrant the conclusion that folk intuitions *in general* should be rejected in thinking about actual causation, however, but rather that insofar as the debates over actual causation concern the ordinary concept of causation, our theories should be restricted to agents.

[5] See Sytsma et al. (2012), Livengood et al. (2017), Sytsma et al. (2019), Livengood and Sytsma (forthcoming), Sytsma (under review a, b), Sytsma and Livengood (under review).

[6] As argued in Livengood and Sytsma (forthcoming), this is not the only motivation for work on actual causation. And we suggest that "philosophers may accept or reject the FAD to differing degrees with respect to different projects." While many researchers aim to provide a description or analysis of the ordinary concept, for which the FAD would seem non-negotiable, others might be better described as engaging in conceptual engineering or in trying to figure out what causation *really* is. For the conceptual engineer, divergence from lay intuitions is to be

Not surprisingly, the bias and responsibility views make the same predictions about most cases that have been investigated in the empirical literature. The central examples involve scenarios where two agents jointly bring about a bad outcome, with one agent violating an explicit injunctive norm while the other does not. For such cases, participants are more likely to treat the agent who violates the norm as the cause of the outcome than the agent who does not violate the norm. And a comparable effect is found when there is just a single agent bringing about the outcome—that agent being treated as more causal in scenarios where she violates a norm than in scenarios where she does not (Livengood et al. 2017). Both the bias view and the responsibility view can readily explain these results, the former holding that they reflect that norm violations bias ordinary causal attributions, the latter that they reflect the appropriate application of a normative concept.

As discussed in Livengood and Sytsma (forthcoming), deciding between the bias and responsibility views poses the challenge of distinguishing between the systematic misapplication of a concept and the correct application of an alternative concept. In that paper, we argue that the bias view would require a self-underminingly large bias to explain the empirical results, and that both charity and simplicity favor the responsibility view. In the present paper I expand on this, offering empirical evidence bearing directly on the dispute. I test a key point of divergence between the two accounts: while both the bias and responsibility views expect that various features of an agent's mental states that are relevant to assessing responsibility for the outcome will impact ordinary causal attributions, the bias view goes further, holding that purely peripheral

expected, although an understanding of the ordinary concept is likely to be useful for this project. For realist metaphysics, the FAD is likely to be reasonable only insofar as it raises worries about having changed the topic and corresponding worries about motivating the topic. Given that the motivation for Rose's debunking argument is tied to the assumption that "fit with folk intuitions of actual causation is taken to serve as an important desideratum in evaluating theories of actual causation" (2017, 1325), I will assume a motivation that implies the FAD for present purposes.

features of the agent—features that are irrelevant to appropriately assessing responsibility, such as the agent's general character—will also impact causal attributions. In contrast, while the responsibility view allows that people will sometimes show bias, we expect that features that are irrelevant to appropriately assessing responsibility should not have a major impact on ordinary causal attributions.

Here is how I will proceed. In Section 1, I briefly lay out the bias and responsibility views. In Section 2, I consider a previous experiment in the literature—the first study in Alicke (1992)—that bears on the role of character in ordinary causal attributions. I explore this case further in Section 3, conducting two follow-up experiments. The results suggest that character does not have an independent effect; rather, in this case participants draw an inference from character to negligence. In Section 4, I investigate the role of character for a different type of scenario, building off the findings of Livengood et al. (2017). The results again suggest that character does not have an independent effect, instead this time working through inferences to the agent's knowledge and desire in performing their action. I then explore the relative roles of knowledge and desire on causal attributions, finding that each has an impact, although what impact they have is complicated by further inferences. I conclude that the present evidence favors the responsibility view over the bias view, casting doubt on Rose's debunking argument.

1. Bias and Responsibility

In this section I discuss two competing accounts of the impact of norms on ordinary causal attributions—the bias view and the responsibility view—focusing on one key difference: the bias view, but not the responsibility view, expects that peripheral features of the agent will also have a notable impact on causal attributions.

The responsibility view holds that the concept of causation that lay people typically employ when they make causal attributions has a normative component.[7] The basic claim is that causal attributions serve to indicate more than that an agent contributed to bringing about an outcome—they also give a normative evaluation of the action akin to saying that the agent is responsible or accountable for the outcome. As such, the responsibility view offers a direct explanation of the effect of norms on ordinary causal attributions: norms play a role because the concept at play is in part normative.[8] One upshot of this view is that the impact of norms does not reflect bias or error; rather, people are correctly applying the relevant concept. Given that the resulting causal attributions often diverge from those given by philosophical accounts of actual causation, the implication is that philosophers are either making a mistake in applying the ordinary concept or else tapping judgments about a different concept, then mistakenly taking that concept to be the ordinary concept at play in causal attributions.

The responsibility view holds that factors beyond just the agent's action and the outcome she contributes to will often matter for causal attributions, since they often matter for responsibility judgments. This includes factors concerning the agent's mental states, such as whether she knew about the norm (or should have known this), whether she knew that her action would contribute to the outcome (or should have known this), whether she desired to violate the

[7] To put this more carefully, "people" should be restricted throughout to native English-speaking adult Americans—the population that most of the participants in the relevant studies have been drawn from (but see Samland and Waldmann 2016 for similar findings using a more geographically diverse sample; see Samland et al. 2016 for findings on children). The extent to which these findings generalize to other English-speakers and, especially, to attributions in other languages remains an interesting open question.

[8] The responsibility view is similar to the *pragmatic view* put forward by Samland and Waldmann (2015, 2016), although they make a subtly different claim. Samland and Waldmann's view is that the causal queries used in the empirical literature are often ambiguous and that pragmatic features of the probes tend to lead participants to interpret the questions as asking about accountability rather than descriptive causation. The responsibility account goes further than this, asserting that the "accountability" usage of the language of causal attribution is the dominant use. In other words, contra Samland and Waldmann, we do not believe that pragmatic features of the experimental setups are shifting participants away from a descriptive reading of "caused" that they are otherwise likely to employ; rather, we propose that these probes are accurately eliciting judgments about the applicability of the ordinary concept of causation at play in causal attributions. See Sytsma et al. (2019) for corpus evidence supporting this claim.

norm, and whether she desired for the outcome to occur. For instance, the reasoning behind a key prediction in Sytsma et al. (2012, 816) included that "the agent could reasonably be expected to know that a bad outcome might result from her behavior." Our expectation was that many factors go into assessing responsibility for an outcome, including most prominently whether it is judged that an agent foresaw, or should have foreseen, that her action would contribute to the outcome and whether she desired to bring about the outcome. Further, a good deal of empirical evidence suggests that such factors impact responsibility attributions and related judgments (e.g., Cushman 2008, Gailey and Falk 2008, Lagnado and Channon 2008, Malle et al. 2014, Young and Saxe 2011, Samland and Waldmann 2016).

In addition, there is some existing evidence that mental state factors matter for ordinary causal attributions. For instance, Samland and Waldmann (2016) test the agent's knowledge of the norm that is violated. In their fourth experiment they use a cover story in which a gardener (Benni) uses a fertilizer he shouldn't have, leading to some plants drying up. In one condition, Benni uses the illicit fertilizer intentionally, knowing the rule against doing so; in a second condition, he uses the illicit fertilizer accidently; in a third condition, he uses the fertilizer intentionally but doesn't know about the rule against doing so; and, in the fourth condition, he is tricked into using the fertilizer. When Samland and Waldmann asked which agent caused the outcome, they found that a large majority of participants selected Benni in the first condition, but that this decreased dramatically in the other three conditions, with only a small minority selecting Benni in the fourth condition.[9] And Kirfel and Lagnado (2017) present evidence that

[9] This study has been successfully replicated by Kominsky and Phillips (2019), who advocate for another view in the literature—the *counterfactual view* (Hitchcock and Knobe 2009, Kominsky et al. 2015, Icard et al. 2017, Kominsky and Phillips 2019). While Samland and Waldman take their results to push against the counterfactual view, Kominsky and Phillips argue that the counterfactual view is also able to explain the findings. In brief, the counterfactual view holds that the effect of norms works through making the counterfactual on which the norm wasn't violated more salient. Kominsky and Phillips then expand on this, suggesting that information about an agent's mental states can affect whether the agent is taken to have violated a norm. Specifically, they focus on

the agents' knowledge of the likely outcome of their action also matters for ordinary causal attributions. In two experiments they found that causal attributions were higher when the agents knew about the likely outcome of their actions than when they did not know about the likely outcome of their actions.

1.2 Bias

The bias view put forward by Alicke and colleagues—otherwise referred to as the culpable control model or blame validation account—sees the process of making causal attributions as one in which the desire to blame biases otherwise descriptive causal judgments. The basic idea is that ordinary causal attributions are often implicitly shaped to rationalize or validate our desires to blame or praise. Thus, Alicke et al. (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"; this has the result that "causal attributions reflect a desire to praise or denigrate those whose actions we applaud or deride." As such, the bias view holds that the impact of norms is the result of people making a mistake in applying the ordinary concept of causation. As Rose (2017, 1327, italics in original) puts it, "on this view, the effect of moral considerations on folk intuitions of actual causation is an *error*, rooted in a motivational bias to blame those who engage in harmful or offensive actions." If we also assume that the judgments of philosophers about causal scenarios are generally free from this bias, then we would have reason to favor philosophical intuitions in analyzing the ordinary concept of causation.

Like the responsibility view, the bias view does not hold that norm violations are the only thing that will matter for ordinary causal attributions. The bias view allows that mental state

whether the agent knew and/or desired to violate the norm. While I will not focus on the counterfactual view in this paper, it is unclear how readily this view can explain the new results reported in Section 4, which focus on whether the agent knew and/or desired to bring about the outcome.

factors like those discussed above might play a role in people's desire to blame or praise an agent, which would then impact their causal attributions. The bias view goes further than this, however, also holding that "peripheral features" of the agent (Alicke et al. 2011)—features that arouse negative/positive evaluations but are not relevant to appropriately assessing the agent's responsibility for the outcome—should impact causal attributions. They hold that such evaluations can occur in response to features directly relevant to the outcome and how it came about, but also in response to "peripheral features of the event such as the actor's or victim's race or character" (674). If peripheral features of the agent, such as her general character, have a notable impact on ordinary causal attributions, it seems this would be best understood in terms of bias, and as a result would favor the bias view over the responsibility view.

2. Motive and Character

Most of the cases tested in the empirical literature manipulate whether or not an agent violates a norm. The first case tested by Alicke (1992), however, is a bit different. The scenario he used involves a driver (John) who violates an injunctive norm (drives 40 mph in a 30-mph zone) and hits another car. The manipulation of interest for present purposes doesn't directly involve this norm violation, however, but John's *motive* for violating it. In one set of conditions John's motive was good (he was speeding "in order to get home in time to hide an anniversary present for his parents that he had left out in the open before they could see it") and in the other set of conditions his motive was bad (he was speeding "in order to get home in time to hide a vial of cocaine he had left out in the open before his parents could see it"). Alicke found that people were more likely to answer that John was the primary cause of an accident when he had a bad motive than when he had a good motive.

Alicke also varied a second contributing cause. Participants were either told that there was an oil spill on the road, that a tree branch was blocking the stop sign, or that the other car ran a stop sign. For present purposes, the first version is most relevant, since the link from the norm violation to the outcome is the clearest. In the oil spill conditions, John "applied his brakes, but was unable to stop as quickly as usual because of some oil that had spilled on the road." The result was that "John hit a car that was coming from the other direction… on the driver's side, causing him multiple lacerations, a broken collar bone, and a fractured arm." In the oil spill conditions, participants were notably more likely to cite John as the primary cause of the accident in a sentence completion task when he had a bad motive (96.6%) than when he had a good motive (65.5%), and a similar gap was seen for responsibility ratings, causal ratings, and the amount of compensation that participants thought the other driver deserved.

At first glance, these results plausibly support the bias view over the responsibility view.[10] John's motive for speeding is arguably informing participants about a peripheral feature of the agent—his general character. Based on just the information provided in the vignettes, in one condition John would seem to be a good son, in the other a bad son. And if such peripheral features have a significant effect on causal attributions, this would be evidence in favor of the

[10] In fact, this is the first piece of evidence that Rose (2017) notes in support of the bias view and the first prong of his debunking argument. He also cites the causal modeling work conducted by Alicke et al. (2011) and our joint work in Sytsma et al. (2012) in responding to the counterfactual view, although interestingly he doesn't note the alternative responsibility view that we first put forward in that paper. The relevant study from Alicke et al. is Study 3, which involves a man named Poole entering Turnbull's house and being shot. The key manipulation for our purposes is whether Poole was described as a dangerous ex-convict who broke into Turnbull's home or a physician who Turnbull's wife had asked to feed her cat while she was away. Alicke et al. found that ratings for both a causal attribution and a blame attribution were higher in the former condition than the latter. While they describe this manipulation in terms of Turnbull's "character," it is rather clear that the description is relevant to assessing Poole's culpability: one is surely more warranted in shooting a dangerous convict who has broken into your house than someone who was invited into the home and is doing a homeowner a favor. As such, these findings do not suggest against the responsibility view. Alicke et al. also detail two statistical analyses suggesting that the best model of the relationship is that blame judgments cause causal judgments. This includes that blame mediates the effects of the manipulation on causal ratings. Again, these findings do not suggest against the responsibility view, which expects the effect of the description of Poole to impact causal attributions through their relevance to assessing responsibility.

bias view. Arguably, neither motive for John's speeding should notably mitigate his responsibility for the outcome, although a story to that effect could perhaps be constructed.

It is not clear that John's motives in Alicke's scenarios can simply be treated as informing on peripheral features of the agent, however. The reason is that John's motive seems relevant for purposes of inferring something about John's skill as a driver, which is not a peripheral feature of the agent, at least with regard to assessing whether he caused a car crash. For instance, rushing home to hide a vial of cocaine that had been left out in the open suggests a serious drug problem, and a serious drug problem raises questions about that person's driving ability. As Malle et al. (2014, 163) put the point: "in real life an agent's goals (and inferred character) may provide preventability information: for example, that the drug-hiding agent was driving faster, was more inattentive, and more careless than the gift-hiding agent, warranting greater causality and blame judgments." One way to assess whether the effect on causal attributions seen in Alicke's driver experiment reflects participants' negative evaluations of the driver's character or whether participants instead draw an inference to his driving ability, is to modify the case so that the driver's character is indicated in a way that does not as clearly suggest an inference to driving ability. This was the goal of my first set of studies.

3. Driver Studies

To better test whether participants' judgments in Alicke's (1992) driver experiment reflect bias or an inference to driving ability, I ran two studies using simplified versions of his vignettes. In the first I varied information pertinent to assessing the driver's character. In the second I varied information directly pertaining to her driving ability.

3.1 Study 1

In my first study, participants were given one of two scenarios describing a driver (Samantha) who hits another vehicle.[11] Samantha was either described in a way intended to generate a positive assessment of her ("Samantha likes to tutor underprivileged children in her spare time") or a negative assessment of her ("Samantha likes to torture cats in her spare time"), but where neither description would seem to warrant a direct inference about her driving ability. In both scenarios Samantha was then described as being involved in a car accident while driving home from work. Participants were told that "Samantha hit the other driver on the driver's side, causing him multiple lacerations, a broken collar bone, and a fractured arm." After reading the vignette, participants were asked to "Complete the following sentence: The primary cause of this accident was ___________." On a second page, participants were asked to rate their agreement with each of two claims—"Samantha was responsible for the accident" and "Samantha caused the accident"— using a 7-point scale anchored at 1 with "totally disagree," at 4 with "neither agree nor disagree," and at 7 with "totally agree." Finally, participants were asked about compensation:

> The law allows between $0 and $100,000 to be awarded to the victim in an automobile collision. In cases such as this, if Samantha were believed to be at fault, the average award would be about $25,000. How much money, if any, do you think should be awarded to the driver that Samantha hit?

The vignette was repeated on the second page and participants were not able to go back to the first page. Responses for Study 1 were collected from 132 participants.[12]

[11] Full vignettes for each study in this paper are given in the supplemental materials.

[12] Participants for each study presented in this paper were recruited through advertising for a free personality test on Google. Prior to the causal scenario, participants answered basic demographic questions. After the causal scenarios they took a 10-item Big Five personality inventory. Participants for each study were restricted to native English-speakers, 16 years of age or older. Participants for Study 1 were 64.4% women with an average age 35.7 years and ranging in age from 16 to 71. Given the higher percentage of woman, I checked for gender effects on participant responses. One-way ANOVAs showed no significant gender effects on responsibility ratings, causal ratings, or compensation.

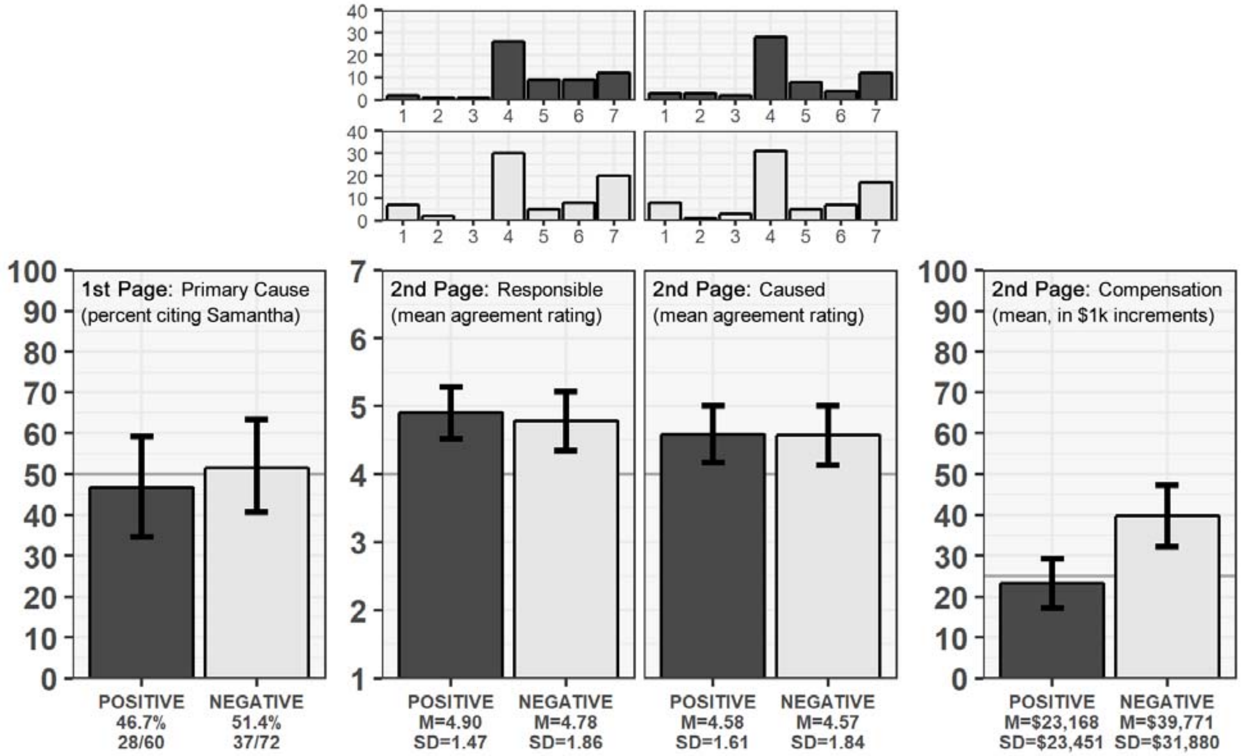

Figure 1: Results of Study 1 showing 95% confidence intervals; histograms of responses shown above.

Results are shown in Figure 1. There was no significant difference between the two conditions in how often participants identified Samantha as the primary cause, in mean agreement with the responsibility claim, or in mean agreement with the causal claim.[13] In other words, whether Samantha was described in a way indicating that she is a good or a bad *person* had no notable effect on participants' causal judgments. As such, it might be worried that the descriptions used simply did not elicit positive/negative evaluations of Samantha from participants, despite their rather extreme nature. Responses to the compensation question indicate that this is not the case, however. While the description of Samantha had no effect on causal

[13] Primary Cause: $\chi^2(1)=0.13$, $p=0.71$. Responsibility Rating: $t(129.69)=0.42$, $p=0.67$, d=0.072; $W=2197.5$, $p=0.86$. Causal Rating: $t(129.72)=0.046$, $p=0.96$, d=0.0080; $W=2160.5$, $p=1$.

judgments, it had a notable effect on how much compensation participants thought the other driver deserved. In the condition where Samantha was described as torturing cats, the average compensation assigned was 71.7% higher than when she was described as tutoring underprivileged children, and the difference was statistically significant.[14] This indicates that participants were in fact forming positive/negative evaluations of Samantha even though these evaluations had no notable effect on their causal attributions.

3.2 Study 2

In my second study, participants were given one of two new variations on the scenarios from Study 1. This time, instead of describing Samantha in ways intended to generate positive or negative assessments of her, I directly stated that she was either a good driver ("Samantha is a very safe and attentive driver") or a bad driver ("Samantha is a very unsafe and inattentive driver"). Samantha was again described as being involved in a car accident while driving home from work and participants were asked the same four questions as in the previous study. Responses were collected from 138 participants.[15]

Results are shown in Figure 2. This time there were significant differences between the two conditions for each of the four questions, with responses being higher when Samantha was described as bad driver compared to when she was described as a good driver.[16] Comparing the two studies, whether Samantha was described in ways indicating that she is a good or a bad *person* had no notable effect on participants' causal judgments, but whether she was said to be a

[14] $t(128.13)=3.44$, $p=0.00078$, d=0.59

[15] 70.3% women with an average age 35.9 years and ranging in age from 16 to 72. One-way ANOVAs showed no significant gender effects.

[16] Primary Cause: $\chi^2(1)=8.38$, $p=0.0038$. Responsibility Rating: $t(134.15)=3.42$, $p=0.00084$, d=0.58; $W=1502.5$, $p=0.00012$. Causal Rating: $t(133.92)=4.10$, $p=7.1\text{e-}5$, d=0.70; $W=1493.5$, $p=0.00010$. Compensation: $t(134.17)=2.02$, $p=0.046$, d=0.34.

good or a bad *driver* had a significant effect. This suggests that the results of Alicke’s (1992) driver experiment are best interpreted in terms of participants inferring something about John’s driving ability from his motives, not as showing an independent effect of character on causal attributions. Further, while the results of our first two studies are in line with the responsibility view, they are at odds with the bias view: the bias view would predict that participants’ evaluations of Samantha in the first study should impact their causal judgments, as they impacted their compensation judgments, but they did not.

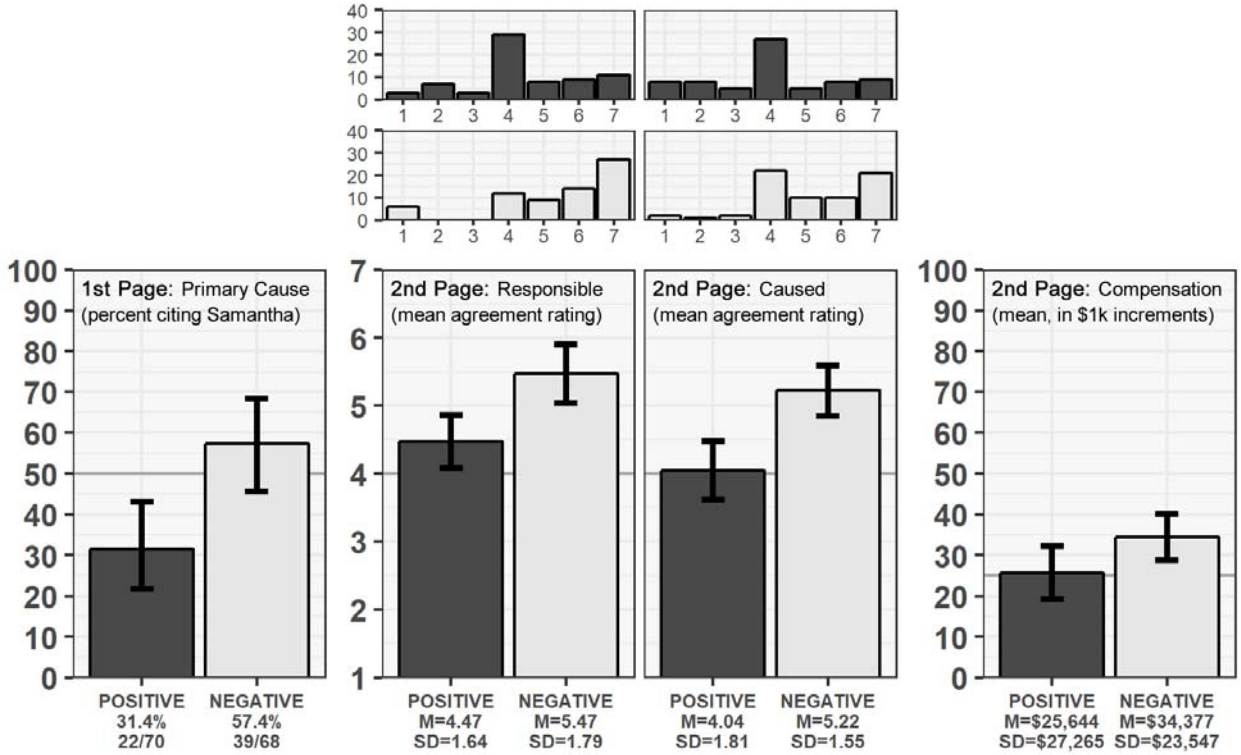

Figure 2: Results of Study 2 showing 95% confidence intervals; histograms of responses shown above.

4. Lauren Alone Studies

Most of the scenarios that have been examined in the recent empirical literature on ordinary causal attributions involve two agents that perform symmetric actions that jointly bring about an outcome. The Driver Case diverges from this insofar as it highlights just one agent, although the other driver is obviously involved as well. In Study 17 in Livengood et al. (2017), we tested a scenario that went a bit further. In the Lauren Alone Case, just one agent was noted in the scenarios and her action was described as being sufficient to bring about a bad outcome.

In these scenarios, Lauren works for a company that uses a mainframe computer that can be accessed from terminals on different floors of its building. Though the company does not know it, the mainframe has become unstable, such that it will crash if anyone logs into it. Lauren logs into the mainframe and the system crashes. Participants were then asked to rate their agreement with a causal attribution—"Lauren caused the system to crash"—using the same 7-point scale used in my previous studies. On a second page, participants were then told that the company learned that the mainframe was unstable and implemented a policy prohibiting employees from logging in. Although Lauren knew about the policy, she again logged into the mainframe and the system crashed. Participants were then asked the same question as on the previous page. We ran five versions of this case, varying the typicality of Lauren's action (either relative to her own past behavior or the behavior of employees more generally). In each case participants tended to deny that Lauren caused the system to crash on the first page and to affirm that she caused the system to crash on the second page.

In this section I report the results of three further studies using the Lauren Alone Case to test the impact of other features of the agent. In Study 3, I vary whether information is added to indicate that Lauren has a bad character and whether information is added to indicate that she lacked knowledge and desire with regard to the outcome. I find that character has an effect when knowledge and desire are unspecified, but that the effect disappears when it is explicitly stated that Lauren does not know that the outcome will occur or desire for it to come about. In Study 4, I then test the relative role of knowledge and desire in causal attributions, finding that both have an independent effect, with knowledge having a notably larger effect than desire. The case where Lauren knows that the outcome will occur if she logs, does not desire for it to occur, but logs in anyway is explored further in Study 5, giving three reasons for Lauren's seemingly contradictory behavior. I find that which explanation is offered has a dramatic impact on ordinary attributions.

4.1 Study 3

To test the role of other factors besides violation of an explicit injunctive norm in the Lauren Alone Case, I varied whether information about Lauren's character was included (*not specified*, *bad person*) and whether Lauren was described as not knowing or desiring that the system would crash (*not specified*, *lacks knowledge and desire*) using a 2x2 between-participants design. To indicate that Lauren was a bad person, the following paragraph was added to the beginning of the non-normed vignette from the first condition in Study 17 from Livengood et al. (2017):

> Lauren lives alone in a secluded area by a small forest. Her favorite pastime is to set traps for animals in the forest. When she catches one, she brings the animal back to her garage where she keeps it alive for several days slowly torturing the animal in ways too gruesome to mention. She has done this with raccoons, deer, rabbits, and even the occasional cat or dog. Because she lives in a secluded area, however, nobody knows that Lauren does this and her colleagues at work generally think that she is a pleasant person.

To indicate that Lauren neither knew that the system would crash nor desired for it to crash, at the end of the vignette participants were told "Lauren did not know that the system would crash, nor did she have any desire for the system to crash." In each condition, participants were given the same follow-up probe as in Livengood et al.'s Study 17 on a second page.[17] Responses were collected from 309 participants.[18]

Results are shown in Figure 3. Focusing on the first page, an ANOVA with whether Lauren was described as a bad person and whether she was described as lacking knowledge and desire as between-participant factors showed main effects for each, as well as a significant interaction.[19] Planned comparisons showed that the mean response was significantly above the neutral point when Lauren was described as a bad person and her knowledge/desire was unspecified.[20] But in the other three conditions the mean response was significantly below the neutral point.[21] Most importantly, while causal ratings were significantly higher when Lauren was described as a bad person than when she was not in the conditions where knowledge and desire was unspecified, there was no notable difference between the conditions when Lauren was described as lacking knowledge and desire.[22] Interestingly, on the second page when the norm was added, a matching ANOVA showed just a significant interaction effect between the two

[17] Although my primary focus in these studies will be on responses to the unnormed vignette on the first page, results will also be shown for the normed vignette on the second page. To further test this set-up, I replicated the first condition from Livengood et al.'s study using a between-participants design: participants either received the probe from the first page or a modified version of the probe from the second-page, removing reference to the system having previously crashed. Responses were collected from 82 participants (70.7% women, average age 36.7 years, ranging from 16 to 99). For the non-normed condition, the mean was significantly below the neutral point (N=41, M=3.12, SD=2.12; $t(40)=2.65$, $p=0.0058$, d=0.41; $V=156.5$, $p=0.0068$) and the mean was not significantly different from that found by Livengood et al. ($t(82.968)=0.99$, $p=0.32$, d=0.20; $W=1377.5$, $p=0.37$). For the normed condition, the mean was significantly above the neutral point (N=41, M=5.63, SD=1.74; $t(40)=6.00$, $p=2.3\text{e}^{-7}$, d=0.94; $V=584.5$, $p=2.6\text{e}^{-5}$) and, again, the mean was not significantly different from that found by Livengood et al. ($t(89.317)=0.21$, $p=0.83$, d=0.042; $W=1245.5$, $p=0.97$).

[18] 59.1% women (one non-binary) with an average age 36.8 years and ranging in age from 16 to 81. One-way ANOVAs showed no significant gender effect for responses on either page.

[19] Character: $F(1, 305)=10.01$, $p=0.0017$, $\eta^2=0.030$; Knowledge/Desire: $F(1, 305)=6.14$, $p=0.014$, $\eta^2=0.018$; Interaction: $F(1, 305)=13.99$, $p=0.00022$, $\eta^2=0.042$.

[20] $t(88)=2.74$, $p=0.0038$, d=0.29; $V=1833.5$, $p=0.0072$

[21] Character and Knowledge/Desire Not Specified: $t(75)=4.15$, $p=4.4\text{e}^{-6}$, d=0.48; $V=487.5$, $p=8.6\text{e}^{-5}$. Character Not Specified and Lacks Knowledge/Desire: $t(71)=2.47$, $p=0.0079$, d=0.29; $V=719.5$, $p=0.0053$. Bad Character and Lacks Knowledge/Desire: $t(71)=3.65$, $p=0.00025$, d=0.43; $V=425$, $p=0.00027$

[22] $t(140.56)=0.61$, $p=0.54$, d=0.10; $W=2708$, $p=0.63$

factors.[23] While the mean response was significantly above the neutral point in each condition, it was lower when Lauren had a bad character and lacked knowledge and desire.[24] It is unclear how best to explain this effect, and replication is needed given that neither view straightforwardly predicts it.

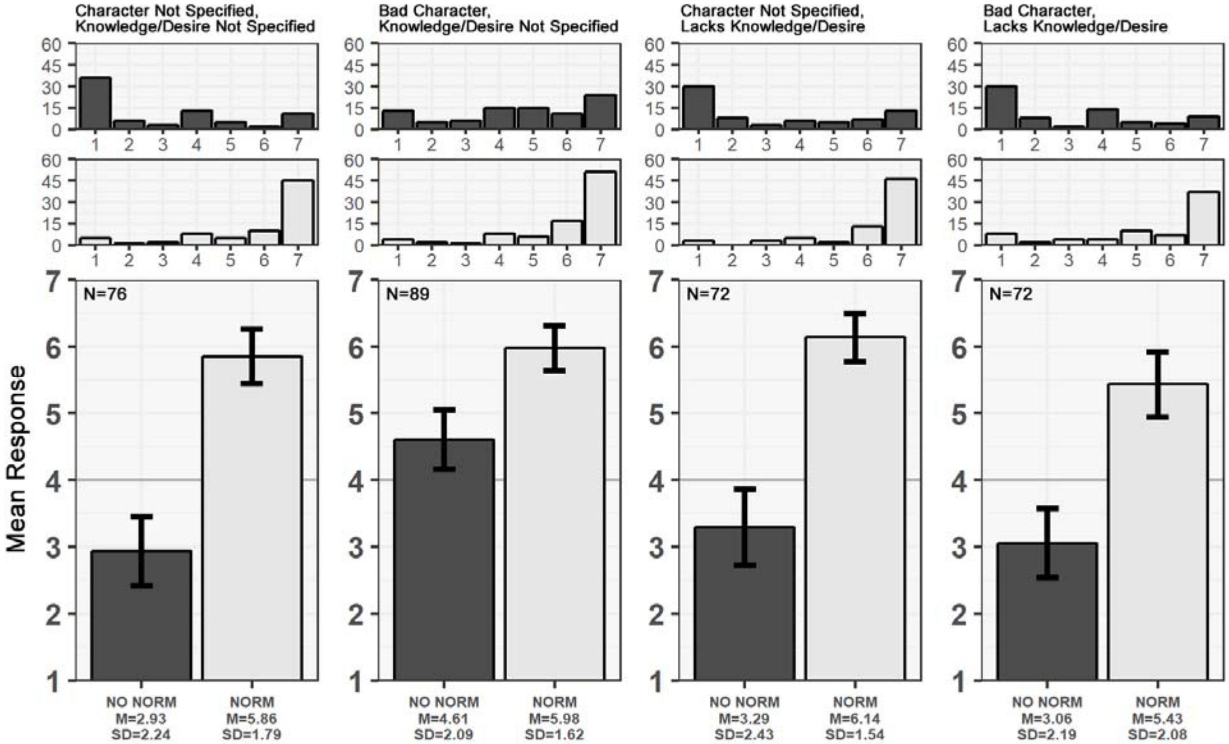

Figure 3: Results of Study 3 showing 95% confidence intervals; histograms of responses shown above.

What we find in Study 3 is that while character has an effect on causal attributions when Lauren's state of knowledge and desire is unspecified, this effect evaporates when participants

[23] Character: $F(1, 305)=1.69$, $p=0.20$, $\eta^2=0.005$; Knowledge/Desire: $F(1, 305)=0.53$, $p=0.47$, $\eta^2=0.002$; Interaction: $F(1, 305)=4.27$, $p=0.040$, $\eta^2=0.014$

[24] Character and Knowledge/Desire Not Specified: $t(75)=9.05$, $p=5.9e^{-14}$, d=1.04; $V=2107.5$, $p=9.2e^{-10}$. Bad Character and Knowledge/Desire Not Specified: $t(88)=11.5$, $p<2.2e^{-16}$, d=1.22; $V=3067$, $p=2.6e^{-12}$. Character Not Specified and Lacks Knowledge/Desire: $t(71)=11.8$, $p<2.2e^{-16}$, d=1.39; $V=2140$, $p=2.4e^{-11}$. Bad Character and Lacks Knowledge/Desire: $t(71)=5.83$, $p=7.5e^{-8}$, d=0.69; $V=1910$, $p=1.5e^{-6}$. Comparing the last two conditions: $t(130.85)=2.32$, $p=0.022$, d=0.39; $W=3044.5$, $p=0.044$.

are told that Lauren neither knew that the outcome would occur nor desired for it to occur. This suggests that the effect of character works via an inference to what Lauren knows and desires: in the absence of information specifying otherwise, people infer from Lauren being a bad person that she likely brought about the bad outcome on purpose. Again, this is in line with the responsibility view but runs counter to the bias view.

4.2 Study 4

Study 3 provided further evidence that character does not have an independent effect on causal attributions, but instead works through further inferences about the agent. In this case, the impact of indicating that Lauren is a bad person was nullified by specifying that she neither knew that her action would lead to the outcome nor desired for the outcome to occur. This suggests that specifying that Lauren knew that the outcome would occur and desired for it to occur would increase causal ratings. This prediction, along with the relative role of knowledge versus desire, were tested in Study 4.

Participants were given one of four versions of the vignette from the first condition of Study 3, varying whether Lauren knew that her action would lead to the outcome (*didn't know*, *knew*) and whether she desired for the outcome to occur (*didn't desire*, *desired*) using a 2x2 between-participants design. This was done by adding a short paragraph to the vignette. To illustrate, in the condition where Lauren both knew and desired, the paragraph reads:

> Lauren knows that the system will crash if she logs into the mainframe. And Lauren has a strong desire for the system to crash.

In each condition, participants were again given the same follow-up probe as in the previous study on a second page. Responses were collected from 327 participants.[25]

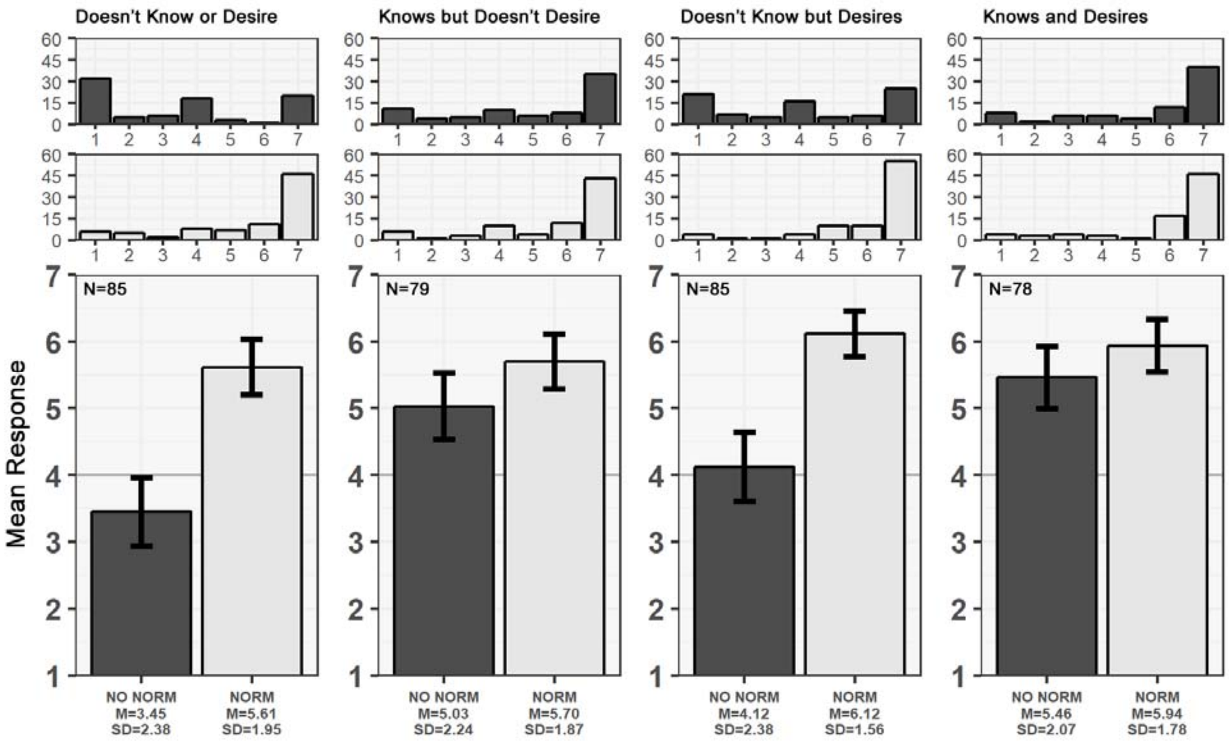

Figure 4: Results of Study 4 showing 95% confidence intervals; histograms of responses shown above.

Results are shown in Figure 4. Focusing on the first page, an ANOVA with knowledge and desire as between-participant factors showed main effects for each, with knowledge having a notably larger effect; no interaction effect was found.[26] Planned comparisons showed that the mean response was significantly below the neutral point when Lauren neither knew that her

[25] 70.0% women (five non-binary) with an average age 46.9 years and ranging in age from 16 to 88. One-way ANOVAs showed no significant gender effect for responses on either page.

[26] Knowledge: $F(1, 323)=33.6$, $p=1.6e\text{-}8$, $\eta^2=0.093$; Desire: $F(1, 323)=4.92$, $p=0.027$, $\eta^2=0.014$; Interaction: $F(1, 323)=0.22$, $p=0.64$, $\eta^2=0.001$.

action would bring about the outcome nor desired for the outcome to occur.[27] When she knew but did not desire, however, the mean response was significantly above the neutral point; further, it was significantly greater than in the first condition.[28] In comparison, when Lauren desired but did not know, the mean response was not significantly different from the neutral point; and, although it was significantly greater than in the first condition, the effect size was small.[29] Finally, the mean response was largest when Lauren both knew and desired, although it was not significantly greater than in the condition where she knew but did not desire.[30] On the second page when the norm was added, a matching ANOVA showed no significant effects, although desire was borderline significant.[31]

Study 4 suggests that, at least for this scenario, causal attributions are sensitive to participants' perception of the agent's knowledge of the likely outcome of her action, and to a lesser extent her desire to bring about that outcome. While the effect of desire was minimal in this study, it is possible that its role is stronger than suggested. One reason to think this is that in the condition where Lauren knew that the system would crash if she logged in and where she did not desire for it to crash, it is unclear why she nonetheless logged in. This might well strike participants as contradictory, leading them to draw a further inference. One possibility is that some participants might have inferred from her behavior that deep-down really Lauren wanted the system to crash. Alternatively, some participants might have inferred that Lauren wasn't paying attention or temporarily forgot that the system would crash, which might lead them to think that she acted negligently. To further test this condition, in Study 5 I gave participants one

[27] $t(84)=2.14$, $p=0.017$, d=0.23; $V=857.5$, $p=0.031$

[28] $t(78)=4.08$, $p=5.5e^{-5}$, d=0.46; $V=1803.5$, $p=0.00011$; $t(161.98)=4.38$, $p=1.1e^{-5}$, d=0.68; $W=2165$, $p=2.6e^{-5}$

[29] $t(84)=0.46$, $p=0.32$, d=0.049; $V=1292$, $p=0.30$; $t(168)=1.84$, $p=0.034$, d=0.28; $W=3041$, $p=0.033$

[30] $t(77)=6.22$, $p=1.2e^{-8}$, d=0.70; $V=2172$, $p=2.8e^{-7}$; $t(168)=1.84$, $p=0.034$, d=0.28; $W=3041$, $p=0.033$

[31] Character: $F(1, 305)=1.69$, $p=0.20$, $\eta^2=0.005$; Knowledge/Desire: $F(1, 305)=0.53$, $p=0.47$, $\eta^2=0.002$; Interaction: $F(1, 305)=4.27$, $p=0.040$, $\eta^2=0.014$.

of three variations on the vignette in which Lauren knew but did not desire, providing a reason for her seemingly contradictory behavior.

4.3 Study 5

Participants were given one of three versions of the vignette from Study 4 in which Lauren knows that her action will lead to the outcome, does not desire for the outcome to occur, but performs the actions anyway, varying the reason given for her behavior. In each case, the reason was added to the beginning of the third paragraph. In the first condition, participants were told that Lauren's boss told her to log in, that Lauren told her boss that the mainframe was unstable, and that her boss told her to log in anyway on threat of being fired. In the second condition, participants were told that Lauren wasn't paying attention and logged in despite knowing that the system would crash if she did so. Finally, in the third condition, participants were told that Lauren's job requires her to log into the mainframe, so she did so despite knowing that it would crash. Participants again answered the same question as in the previous study using the same scale, and they were given the same follow-up probe on a second page. Responses were collected from 204 participants.[32]

Results are shown in Figure 5. One-way ANOVAs with condition as a between-participant factor showed a significant main effect for responses on the first page, but not the second.[33] Planned comparisons showed that the mean response on the first page was not significantly different from the neutral point when Lauren's boss ordered her to log in, but was significantly above the neutral point when she forgot or when she logged in because her job

[32] Participants were 52.5% women with an average age 36.2 years and ranging in age from 16 to 87. One-way ANOVAs showed no significant gender effect for responses on either page.

[33] First Page: $F(2, 200)=10.2$, $p=6.2e^{-5}$, $\eta^2=0.092$; Second Page: $F(2, 200)=0.48$, $p=0.62$, $\eta^2=0.005$.

required it.[34] Most importantly, the mean response was significantly higher when Lauren either forgot or logged in because her job required it, compared to when her boss ordered her to log in.[35] And mean responses for these two conditions were not significantly different from the corresponding condition in Study 4 when no reason was given for why Lauren logged in.[36]

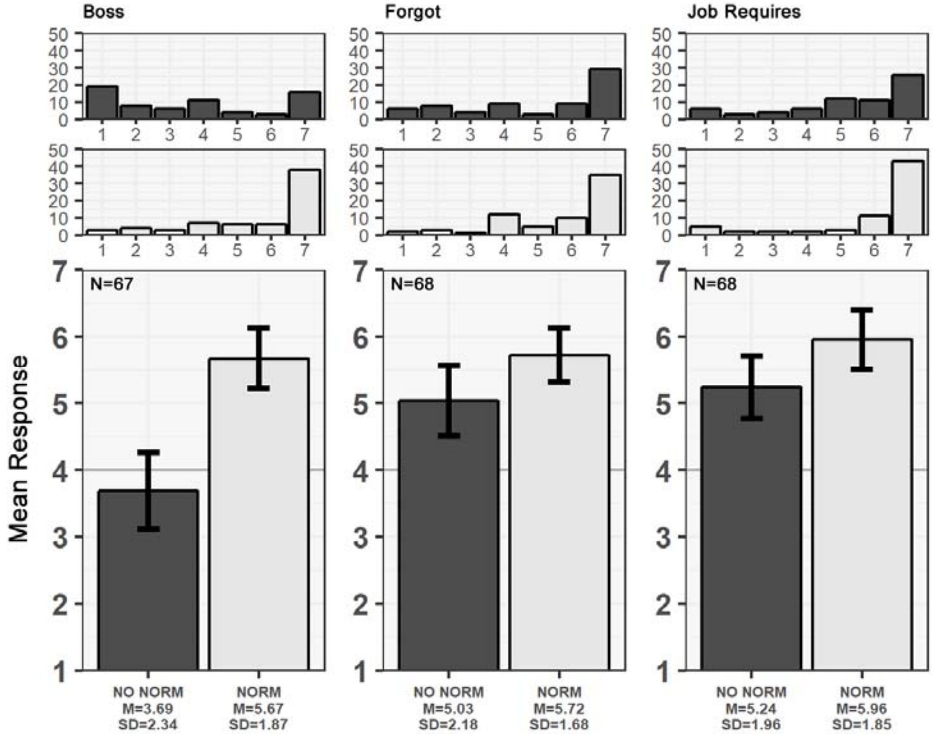

Figure 5: Results of Study 5 showing 95% confidence intervals; histograms of responses shown above.

The results of Study 5 suggest that Lauren's reason for logging in matters when she knows what will happen but does not desire that outcome. Plausibly, desire might be thought of as providing a reason for Lauren to log in (e.g., she logged in because she wanted the outcome to

[34] Boss: *t*(66)=1.09, *p*=0.14, d=0.13; *V*=694, *p*=0.19. Forgot: *t*(67)=3.90, *p*=0.00011, d=0.47; *V*=1374, *p*=7.3e^{-5}. Job Required: *t*(67)=5.21, *p*=9.9e^{-7}, d=0.63; *V*=1569.5, *p*=1.1e^{-5}.

[35] Forgot versus Boss: *t*(131.98)=3.45, *p*=0.00038, d=0.59; *W*=1554, *p*=0.00055. Job Required versus Boss: *t*(128.19)=4.17, *p*=2.8e^{-5}, d=0.72; *W*=1448.5, *p*=9.7e^{-5}.

[36] Forgot versus Not Specified: *t*(142.77)=0.011, *p*=0.99, d=0.0019; *W*=2697.5, *p*=0.96; two-tailed. Job Required versus Not Specified: *t*(144.96)=0.61, *p*=0.54, d=0.099; *W*=2645, *p*=0.87.

occur), but it is not the only reason that she could have. The difference in findings for the three reasons tested in Study 5 can reasonably be explained in terms of the responsibility view. In the condition where Lauren logged in because her boss ordered her to, one might think that her responsibility for the outcome is mitigated: she did what she could to inform her boss about what would happen and only logged in under duress; as such, the responsibility would seem to fall on Lauren's boss, not Lauren. In contrast, when Lauren logged in because she wasn't paying attention, one might think that she acted negligently; and, when she logged in because her job required it, one might think that she should have told the company about the issue with the mainframe, rather than simply ignoring it.

Taking the results of these three studies together, one methodological upshot for work on causal attributions is that researchers should carefully consider what inferences participants might draw about the agents' mental states and reasons for acting, as these can have a notable effect on their causal attributions. Given the impact that such inferences can have, and that the typical scenarios in the literature involve social situations and norm violations that seem likely to elicit such inferences, testing the scenarios without adequately controlling for what participants think the agents know, desire, or the other reasons they might have for acting risks conflating the impact of norms with other factors and could generate misleading results.

5. Conclusion

There is a growing consensus that norms matter for ordinary causal attributions, although there is a great deal of disagreement about why they matter. Each of the main accounts in the literature allows that factors about the agent who violates the norm in a causal scenario, beyond the mere fact that she violates the norm, are likely to impact causal attributions. There is disagreement, however, about exactly which factors are likely to matter. Alicke's bias view holds that not only do features of the agent's mental states matter, such as her knowledge and desires concerning the norm and the outcome, but also peripheral features of the agent whose impact could only reasonably be explained in terms of bias. In contrast, our responsibility view holds that the impact of norms does not reflect bias, but rather that ordinary causal attributions issue from the appropriate application of a concept with a normative component. As such, we predict that while judgments about the agent's mental states that are relevant to adjudicating responsibility will matter, peripheral features of the agent will only matter insofar as they warrant an inference to other features of the agent that are relevant.

In line with the responsibility view and against the bias view, the results of the studies presented in this paper suggest that information relevant to assessing an agent's character matters but only when it warrants an inference to a non-peripheral feature, such as the agent's negligence in the situation or her knowledge and desire with regard to the outcome. Further, the results indicate that information about an agent's knowledge and desire both impact ordinary causal attributions in the scenario tested. This raises an important methodological issue for empirical work on ordinary causal attributions: researchers need to carefully consider and control for the inferences that participants might draw concerning the agents' mental states and motivations.

Further, in presenting evidence that runs counter to the bias view, the present research also suggests against a recent debunking explanation offered for ordinary causal judgments about agents based on this view. Rose (2017, 1352) argues that the "discussion over actual causation should be liberated from any demanded conformity with the folk intuitions" and that "in the dispute over actual causation, folk intuitions deserve to be rejected." But this argument assumes that philosophers working on actual causation are analyzing the ordinary concept. As such, if ordinary causal attributions reflect the appropriate application of the ordinary concept of causation, at least with regard to agents, they cannot be reasonably ignored in pursuing this project. And the present evidence points toward this being the case, rather than the impact of norms on ordinary causal attributions reflecting systematic bias.

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. and D. Rose (2010). "Culpable control or moral concepts?" *Behavioral and Brain Sciences*, 33: 330–331.

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

Cushman, F. (2008). "Crime and Punishment: Distinguishing the Roles of Causal and Intentional Analyses in Moral Judgment." *Cognition*, 108: 353–380.

Danks, D., D. Rose, and E. Machery (2014). "Demoralizing Causation." *Philosophical Studies*, 171: 251–277.

Gailey, J. and R. Falk (2008). "Attribution of Responsibility as a Multidimensional Concept." *Sociological Spectrum*, 28: 659–680.

Hilton, D. and B. Slugoski (1986). "Knowledge-based causal attribution: The abnormal conditions focus model." *Psychological Review*, 93: 75–88.

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

Icard, T., J. Kominsky, and J. Knobe (2017). "Normality and Actual Causal Strength." *Cognition*, 161: 80–93.

Kirfel, L. and D. Lagnado (2017). "'Oops, I did it Again': The Impact of Frequency on Causal Judgements." *Proceedings of the 39th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Knobe, J. and B. Fraser (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., J. Phillips, T. Gerstenberg, D. Lagnado, and J. Knobe (2015). "Causal superseding." *Cognition*, 137: 196–209.

Kominsky, J. and J. Phillips (2019). "Immoral Professors and Malfunctioning Tools: Counterfactual Relevance Accounts Explain the Effect of Norm Violations on Causal Selection." *Cognitive Science*, 43(11): e12792.

Lagnado, D. and S. Channon (2008). "Judgments of Cause and Blame: The Effects of Intentionality and Foreseeability." *Cognition*, 108: 754–770.

Livengood, J., J. Sytsma, and D. Rose (2017). "Following the FAD: Folk attributions and theories of actual causation." *Review of Philosophy and Psychology*, 8(2): 274–294.

Livengood, J. and J. Sytsma (forthcoming). "Actual causation and compositionality." *Philosophy of Science*.

Malle, B., S. Guglielmo, and A. Monroe (2014). "A Theory of Blame." *Psychological Inquiry*, 25: 147–186.

Reuter, K., L. Kirfel, R. van Riel, and L. Barlassina (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., 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.

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.

Sytsma, J. (under review a). “Structure and Norms: Investigating the Pattern of Effects for Causal Attributions.” http://philsci-archive.pitt.edu/16626/

Sytsma, J. (under review b). “The Effects of Single versus Joint Evaluations on Causal Attributions.” http://philsci-archive.pitt.edu/16678/

Sytsma, J. and J. Livengood (under review). “Causal Attributions and the Trolley Problem.” http://philsci-archive.pitt.edu/16200/

Sytsma, J., J. Livengood, and D. Rose (2012). “Two types of typicality: Rethinking the role of statistical typicality in ordinary causal attributions.” *Studies in History and Philosophy of Biological and Biomedical Sciences*, 43: 814–820.

Sytsma, Justin, Roland Bluhm, Pascale Willemsen, and Kevin Reuter (2019). “Causal Attributions and Corpus Analysis.” In E. Fischer and M. Curtis (eds.), *Methodological Advances in Experimental Philosophy*, London: Bloomsbury Press.

Young, L. and R. Saxe (2011). “When Ignorance is No Excuse: Different Roles for Intent Across Moral Domains.” *Cognition*, 120: 202–214.