Crossed Wires: Blaming Artifacts for Bad Outcomes[1]

Justin Sytsma

Philosophers and psychologists often assume that responsibility and blame only apply to certain agents. Sometimes this is nuanced by claiming that there are multiple ordinary concepts of blame and responsibility, with one set being purely descriptive while the other is distinctively moral, and with the latter applying just to certain agents. But do our ordinary concepts of responsibility and blame reflect these assumptions? In this paper, I investigate one recent debate where these assumptions have been applied—the back-and-forth over how to explain the impact of norms on ordinary causal attributions. I investigate one prominent case where it has been found that norms matter for causal attributions, but where it is claimed that responsibility and blame do not apply because the case involves artifacts. Across five studies (total N=1,393) more carefully investigating Hitchcock and Knobe's (2009) Machine Case, I find that the same norm effect found for causal attributions is found for responsibility and blame attributions, with participants tending to ascribe both to a norm-violating artifact. Further, the evidence suggests that participants do so because they are applying broadly normative, but not distinctively moral, concepts.

Legend has it that the first "person" exiled to Siberia was not a person at all, but a church bell. The bell was rung in honor of the passing of Ivan the Terrible's son Dmitry, who died under mysterious circumstances in Uglich. Upon learning of the death, the people of the town rose up, destroying property and killing a high-ranking official. After the uprising was quelled, there came reprisal, including punishment of the bell. Among other things, the bell was publicly flogged and its "tongue" was cut out so that it could never be rung again. And, as if that was not enough, the bell was then exiled to Tobosk (Haywood 2010).

There is no doubt something comical about a church bell being publicly flogged. The bell cannot learn from the punishment or otherwise alter its behavior, and the bell cannot feel pain or otherwise suffer. A church bell simply does not seem to be a fitting target for punishment, whether aimed at rehabilitation or retribution. And yet, if you are anything like me, you've probably gotten angry at a malfunctioning device before. Perhaps you've yelled at your computer

[1] I want to thank Paul Henne, Joshua Knobe, Shaun Nichols, and John Schwenkler for helpful comments on a previous draft of this paper.

when it crashed with a document unsaved or kicked your car when it broke down and left you stranded on the side of the road. Rational or not, it seems that we do blame artifacts, sometimes even going so far as to act on those judgments.

Despite this, philosophical discussions often assume that responsibility and blame only apply to agents, and typically just to people who are sufficiently reason-responsive. This assumption generally appears to be descriptive, holding that ordinary responsibility and blame attributions are typically only applied to certain agents. I'll refer to this as the *agent assumption*. In this paper I challenge the agent assumption, focusing on its application to one recent debate—disagreements about how best to explain the impact of normative considerations on ordinary causal attributions (claims of the form "X caused Y"). This dispute concerns whether norms merely impact causal attributions *indirectly* or whether they play a more immediate role, with our normative evaluations *directly* influencing our causal attributions. Accepting the agent assumption, one compelling piece of evidence for indirect accounts is that the impact of norms on causal attributions is also found in cases involving non-agents, such as the judgments about wires seen in Hitchcock and Knobe's (2009) Machine Case. In this paper, however, I present evidence that the agent assumption does not hold.

Across five studies further investigating the Machine Case, I find that the impact of norms found for causal attributions is also found for responsibility and blame attributions, and that each of these judgments is best explained by people's broadly normative, but not distinctively moral, evaluations of the wires. These results favor direct accounts of the impact of norms on causal attributions over indirect accounts, and they favor our responsibility view over other direct accounts. Here is how I will proceed. In Section 1, I discuss the agency assumption. In Section 2, I show how this assumption factors into recent debates concerning the effect of

norms on causal attributions. And, in Section 3, I present the results of the new studies examining the Machine Case.

1. The Agency Assumption

My concern in this paper is with the agent assumption—the claim that ordinary concepts of responsibility and blame are restricted to certain agents—and its application to debates concerning ordinary causal attributions. While the agent assumption, so formulated, makes a descriptive claim about ordinary concepts, philosophical work on responsibility and blame often has both descriptive and prescriptive components, with authors advancing claims about what these concepts actually look like *and* about what they should look like. That said, the agent assumption generally appears to be assumed of the ordinary concepts, not offered as a proposed revision to those concepts. But even if the appearances are misleading, the accuracy of the agent assumption is important if we are to assess how radically revisionist various philosophical accounts of blame and responsibility are.

Consider Eshleman's (2016) *Stanford Encyclopedia of Philosophy* entry on "Moral Responsibility." He opens by laying out the target concept, and in doing so takes it to be restricted to people and related to concepts like blame:

> When a person performs or fails to perform a morally significant action, we sometimes think that a particular kind of response is warranted. Praise and blame are perhaps the most obvious forms this reaction might take.... Thus, to be morally responsible for something, say an action, is to be worthy of a particular kind of reaction—praise, blame, or something akin to these—for having performed it. (1)

That the target concept is thought to only apply to persons is made clear by a distinction that is then drawn between the concept of moral responsibility and "some others commonly referred to through use of the terms 'responsibility' or 'responsible'" (1). Eshleman continues:

> To illustrate, we might say that higher than normal rainfall in the spring is responsible for an increase in the amount of vegetation.... In [this] case, we mean to identify a causal connection between the earlier amount of rain and later increased vegetation.... Although [this concept is] connected with the concept of moral responsibility discussed here, [it is] not the same, for [in this case we are not] directly concerned about whether it would be appropriate to react to some candidate (here, the rainfall...) with something like praise or blame. (1-2)

To make the reasoning explicit, what Eshleman suggests is that because blame does not apply to non-agents, when we say that a non-agent is responsible for some outcome, we are not using this term in the target normative sense of *moral responsibility*, but instead in the descriptive sense of *causal responsibility*. In other words, taking this discussion to be descriptive, Eshleman employs the agent assumption for blame judgments to carve out a concept of responsibility to which the agent assumption applies.

Interestingly, Tognazzini and Coates's (2016) *Stanford Encyclopedia of Philosophy* entry on "Blame" draws the same sort of distinction between a normative and a descriptive concept, now with regard to blame:

> To begin, note that almost all philosophical discussions of blame ignore (or mention only to set aside) the form of blame sometimes characterized as causal or explanatory responsibility (Kenner 1967; Hart 1968; Beardsley 1969). It is this notion of blame that is at stake when we say that Hurricane Hugo is to blame for the destruction of Charleston's harbor, or that the cat is to blame for knocking over the vase.... Nevertheless, in this entry the focus will be on blame as a response to moral agents on the basis of their wrong, bad, or otherwise objectionable actions or characters. (2-3)

Obviously, there is a bit of tension here. The restriction to (certain) agents for the relevant sense of responsibility is linked by Eshleman to appropriate reactions of praise or blame, but a similar restriction to (certain) agents is asserted by Tognazzini and Coates for blame, now pointing to the agents' "wrong, bad, or otherwise objectionable actions or characters." But this just pushes the issue back a further step. We might now wonder whether people sometimes treat non-persons or non-agents as doing something wrong, bad, or otherwise objectionable.

There is also a good deal of interest in attributions of responsibility and blame in the psychology literature, often taking these to be specifically moral concepts, and now typically with an explicit focus on ordinary attributions. But, again, the agent assumption is common. To give but a few examples, in his widely cited volume on responsibility and blame attributions, Shaver (1985, 66) operates from a working definition of responsibility that is restricted to certain agents—"a judgment made about the moral accountability of a person of normal capacities." He has a similar starting point for blame attributions, writing for instance that "questions about blameworthiness arise only when at least one of the causal elements participating in the production of the effect for which blame is to be assigned is a human action" (162). And more recent work has followed suit. For example, Malle et al. (2014, 148, italics in original) write that blame "is a judgment directed at a *person* who has caused or done something norm violating."[2] Similarly, they hold that "for blame to emerge from the detection of a negative event, the perceiver must establish that an *agent* caused the event" (153).

1.1 Injunctive versus Moral

The distinction we've just seen drawn between causal responsibility or blame and moral responsibility or blame suggests that the former concepts are purely descriptive, while the latter are specifically moral. By purely descriptive, here, I mean that whether the entity violated a norm is not directly relevant to whether causal responsibility or blame apply; what matters is just the causal connection. But it is not clear that the ordinary usage of "responsible" and "blame" follow this philosophical division between concepts.

Alternatively, we might note that specifically moral norms fall within the broader class of *injunctive norms*. Injunctive norms include both prescriptive norms (what should be done) and

[2] Here one wonders what to make of attempts to blame the dog.

proscriptive norms (what should not be done), and while they include specifically moral norms, they are broader than this, covering conventions and etiquette norms, rules and laws, and norms covering how designed systems are supposed to behave (norms of proper functioning). Following this we can draw a further distinction within concepts of responsibility and blame, distinguishing between *normative* concepts and distinctively *moral* concepts. Here, the latter would only apply to moral agents, while the former might apply more broadly. And the former includes the latter: being morally responsible for an outcome is one way of being normatively responsible for that outcome. My contention is that ordinary responsibility attributions and blame attributions are normative, but not necessary moral.[3] And I contend that the agent assumption does not hold for such normative attributions: normative responsibility and normative blame are often attributed to non-agents.

2. Causal Attributions

One place the agent assumption has been employed is in recent debates concerning the effect of normative considerations on ordinary causal attributions. In study after study researchers have found that people are more likely to treat an agent as the cause of a bad outcome when that agent violates an injunctive norm.[4] This is most often demonstrated by asking participants about a scenario in which two agents perform symmetric actions, jointly bringing about a bad outcome,

[3] In fact, this appears to fit with ordinary usage, where entities are frequently said to be responsible for an outcome, but seldom specifically said to be morally responsible for an outcome. To illustrate, a search of the non-academic portions of the Corpus of Contemporary American English (COCA) indicates that while "responsible for" is used frequently (35,917 occurrences), "morally responsible for" is used at most infrequently (0 occurrences). And similarly for "responsible" (55,959 occurrences) and "morally responsible" (108 occurrences). Likewise for "to blame" (12,226) and "morally to blame" (0), "blame" (45,813) and "moral blame" (0), and "blameworthy" (80) and "morally blameworthy" (0).

[4] See, for example, Alicke (1992), Knobe and Fraser (2008), Hitchcock and Knobe (2009), Sytsma et al. (2012), Kominsky et al. (2015), Livengood et al. (2017), Henne et al. (2017), and Kominsky and Phillips (2019) among many others.

but where one agent violates an injunctive norm while the other does not. To illustrate, consider the Pen Case presented by Knobe and Fraser (2009):

> The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take the pens, but faculty members are supposed to buy their own.
>
> The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist has repeatedly e-mailed them reminders that only administrative assistants are allowed to take the pens.
>
> On Monday morning, one of the administrative assistants encounters Professor Smith walking past the receptionist's desk. Both take pens. Later that day, the receptionist needs to take an important message… but she has a problem. There are no pens left on her desk.

Knobe and Fraser found that people tend to agree that the norm-violating agent (Professor Smith) caused the problem and tend to disagree that the norm-conforming agent (the administrative assistant) caused the problem, despite the fact that the two agents did the same thing (both took pens). Call this type of comparative effect of norm-violation/norm-conformity the *norm effect*.

How are we to explain the norm effect for causal attributions? The answer is not obvious, especially given that many philosophers working on causation have taken it to be a purely descriptive matter, such that injunctive norms should not be directly relevant (see Livengood et al. 2017 for discussion). Hitchcock and Knobe (2009) offer an ingenious explanation that preserves the purported descriptive nature of ordinary causal attributions. Their *counterfactual view* contends that norm violations impact causal judgments because they play a role in which counterfactuals people find salient or relevant, and hence which counterfactuals they are most likely to consider.[5] According to this view, while norms are directly relevant to concepts like blame and responsibility, they are only indirectly relevant to casual judgments, with their impact running through the counterfactuals that we consider. Against such *indirect accounts* are

[5] This type of view has been further developed in a number of papers, including Halpern and Hitchcock (2015), Kominsky et al. (2015), Icard et al. (2017), and Kominsky and Phillips (2019).

accounts that contend that normative judgments are directly involved. These *direct accounts* come in many flavors, as discussed below, but each sees the relevant causal judgments as being more intimately tied to normative assessments than indirect accounts allow.

In one way or another, each of the direct accounts calls on normative assessments in a way that indirect accounts do not. As Hitchcock and Knobe (2009, 602-603) put it, direct accounts "rely on a type of moral judgment that plays no role at all in our preferred account," although direct accounts aren't necessarily committed to the relevant norms being specifically moral rather than more broadly injunctive. Following on this, Hitchcock and Knobe propose to give a case where there is a norm violation but no blame attributions:

> In cases of this latter type, the alternative explanations suggest that moral considerations should have no impact on people's causal judgments (because of the absence of blame) while our own hypothesis suggests that the impact of normative considerations should remain unchanged (because people still see that a norm has been violated). (603)

Hitchcock and Knobe offer the Machine Case as such a scenario, making the agent assumption and claiming that in a scenario where there are no agents there will be no blame to be assigned, and similarly, one assumes, no responsibility to be attributed.[6]

The Machine Case is similar to the standard cases in the literature, such as the Pen Case seen above, but with the agents being replaced by artifacts. The scenario reads as follows:

> A machine is set up in such a way that it will short circuit if both the black wire and the red wire touch the battery at the same time. The machine will not short circuit if just one of these wires touches the battery. The black wire is designated as the one that is supposed to touch the battery, while the red wire is supposed to remain in some other part of the machine.
>
> One day, the black wire and the red wire both end up touching the battery at the same time. There is a short circuit.

[6] Hitchcock and Knobe also present a second case in which they contend that blame does not apply—a case involving a good, rather than a bad, outcome. But, recent evidence shows that the norm effect is sometimes *reversed* for cases involving good outcomes, including a variation on the case presented by Hitchcock and Knobe (Schwenkler and Sytsma, ms).

Participants were then asked to rate their agreement with one of two causal attributions on a 1 ("disagree") to 7 ("agree") scale:

> The fact that the red wire touched the battery caused the machine to short circuit.
> The fact that the black wire touched the battery caused the machine to short circuit.

As in previous studies involving two agents performing symmetric actions, in this case participants were much more likely to agree that the norm-violating red wire caused the outcome (M=4.9) than the norm-conforming black wire (M=2.7). Thus, we find the norm effect for a case involving non-agents.[7]

Hitchcock and Knobe take the results for the Machine Case to demonstrate that the norm effect can occur independently of responsibility and blame judgments, holding that this is "a case of norm violation without blameworthiness" (605). This is not something they tested, however; rather the conclusions rests on the agent assumption. Hitchcock and Knobe take the Machine Case to involve a norm violation without blameworthiness exactly because they hold that people do not blame non-agents. Although Hitchcock and Knobe do not note the distinction between causal and moral concepts of responsibility or blame, doing so they might allow that the norm effect will occur for these attributions, but contend that this would not be evidence that the relevant normative judgments are being applied to non-agents. Following the discussion in Section 1, this is somewhat more plausible for responsibility attributions and somewhat less plausible for blame attributions. And it seems still less plausible for other normative judgments, such as fault or punishment attributions. As such, we can formulate a stronger and a weaker prediction concerning the Machine Case:

> **[HK-1]** The norm effect will not occur for responsibility attributions.
> **[HK-2]** The norm effect will not occur for blame, fault, or punishment attributions.

[7] Other cases in the literature have also investigated non-agents, often with agents also being directly involved (e.g., Livengood et al. 2017, Kominsky and Phillips 2019). See Kominsky and Phillips for a recent version of the challenge to direct accounts offered by Hitchcock and Knobe.

If these predictions are accurate, then Hitchcock and Knobe's results would seem to provide compelling evidence against direct accounts, since such accounts tie the norm effect to normative assessments in one way or another.

2.1 Three Responses, Three Direct Accounts

There are at least three ways for advocates of direct accounts to respond to Hitchcock and Knobe's objection, with each being associated most closely with one of the three main types of direct account in the literature, although the responses are not necessarily specific to a given account. The first type of response is to argue that the results for the Machine Case are explained by something other than the fact that one wire violated a norm. Call this the *non-normative response*. Such a response has been offered by Samland and Waldmann (2016) who contend that the Machine Case vignette suggests that the black wire's position is fixed—that it is always touching the battery—while the red wire's position changes. As such, it might simply be that people's causal ratings are picking up on this and treating the red wire as the cause for a reason that has nothing to do with normative considerations. The non-normative response predicts that if we were to change the Machine Case vignette so that both wires move and come to touch the battery at the same time, the (supposed) norm effect for causal attributions should disappear:

[SM] No norm effect for causal attributions when the Machine Case is revised.

Samland and Waldmann offer the non-normative response in defense of one type of direct account—their *pragmatic view*. The pragmatic view explains the impact of injunctive norms on causal attributions by contending that the questions used in the empirical literature demonstrating the norm effect are ambiguous, such that pragmatic features of the probes tend to

lead participants to interpret the questions as asking about normative concepts like accountability or responsibility, rather than the dominant, descriptive concept of causation that is at issue.

The second type of response is to argue that the Machine Case does not provide an example of a norm violation in the absence of blameworthiness because people treat the wires as if they were agents. Call this the *agentive response*. The work of Rose (2017) suggests a response of this type. He notes evidence suggesting that primitive teleological considerations play a role not just in our behavior toward agents, but also non-agents, and he takes this to reflect that we are promiscuous with regard to the entities we treat as agents. Applying this to the Machine Case, the norm effect might be found because participants treat the wires as agents, and so allow that they could be responsible or to blame while still abiding by the agent assumption.

Rose offers a two-pronged debunking explanation of ordinary causal attributions. While one prong concerns primitive teleological considerations, the other corresponds with a second type of direct account—the *bias view*. The bias view contends that the norm effect for causal attributions reflects a general error, with these attributions being biased by people's desire to blame or to praise. The basic idea is that our desires to blame or praise implicitly shape our causal attributions, bringing them in line with our prior evaluations.

The third type of response is to argue that the agent assumption does not hold: participants do not treat the wires as agents, but nonetheless take them to be suitable targets for blame and responsibility. Following the suggestion laid out in the previous section, I predict that by and large people are inclined to judge that the red wire is to blame for and is responsible for the short circuit in the Machine Case, and that they are so inclined because the red wire violated an injunctive norm (a norm of proper functioning). In other words, I contend that the ordinary concepts of blame and responsibility are not specifically moral concepts but are instead more

broadly normative. And while moral concepts may be restricted to moral agents, I predict that normative concepts are applied more broadly. Call this the *normative response*.

Both the agentive and normative responses hold that the norm effects in the Machine Case is truly a *norm* effect. As such, they make the following prediction:

> **[Norm]** Norm effects will be found for normative evaluations and these evaluations will mediate the norm effects for relevant attributions.

While both responses predict **[Norm]**, the normative response goes further, holding that the norm effect in the Machine Case is driven by normative, but not distinctively moral, evaluations; further, it predicts that the norm effect will be notably weaker for distinctively moral judgments, with participants tending to disagree with claims like "the red wire deserves to be punished" and "the red wire did something morally wrong":

> **[NE-1]** Norm effect is primarily mediated by non-moral normative evaluations.
> **[NE-2]** People will tend to disagree with punishment attributions and moral evaluations.

The agentive response, by contrast, does not specifically draw this distinction, and insofar as it holds that the wires are being treated as agents, would expect punishment attributions and moral evaluations to be similar to other normative judgments. In other words, if predictions **[NE-1]** and **[NE-2]** hold, this would suggest in favor of the normative response over the agentive response.

While each of the three types of response I've noted are compatible with each of the three main direct views in the literature, the normative response fits most naturally with our account of the impact of norms on ordinary causal attributions—the *responsibility view*—which focuses on injunctive norms rather than specifically moral norms.[8] Briefly, the responsibility view holds that our normative (but not necessarily moral) evaluations are part of the content considered when applying the ordinary concept of causation at play in causal attributions. And, unlike the previous

[8] See Sytsma et al. (2012), Livengood et al. (2017), Sytsma et al. (2019), Livengood and Sytsma (2020), Sytsma and Livengood (ms), Sytsma (forthcoming).

views, we hold that this is not a mistake: when people judge that a causal attribution is more applicable to an agent who violates an injunctive norm, for example, we contend that they are not misapplying a purely descriptive concept but are correctly applying a concept with a normative component. Thus, we hold that causal attributions typically serve to indicate something more than that someone or something contributed to the outcome or brought about the outcome; they also express a normative evaluation that is roughly akin to saying that the entity is responsible for that outcome. And, in fact, for cases like the Pen Case I've previously found that causal attributions and responsibility attributions are remarkably similar (Sytsma forthcoming). I predict that the same will hold for the Machine Case:

[Responsibility] Causal and responsibility attributions will show a close correspondence.

As discussed in detail in Sytsma (forthcoming), such a close correspondence is problematic for the other views discussed in this paper. First, a suitably close correspondence between causal attributions and responsibility attributions suggests in favor of a common explanation, but an *indirect* account of the impact of norms on *normative attributions* is less plausible than a *direct* account of the impact of norms on *causal attributions*. As such, if **[Responsibility]** holds it would favor direct accounts. Second, the pragmatic and bias views both treat the impact of norms on causal attributions as a type of error, with people either tending to read causal attributions in these experiments as responsibility attributions or their desire to blame tending to bias their causal attributions. While these views would predict *some* correspondence between causal attributions and responsibility attributions, they would not expect them to be overly similar: an especially close correspondence would suggest an implausibly strong error.

3. New Studies

In this section I present the results of a series of five studies further exploring the Machine Case. Study 1 focuses on prediction **[HK-1]**, while Study 2 extends this to **[HK-2]**. Study 3 tests **[SM]**. Finally, Study 4 targets **[Norm]**, while Study 5 tests **[NE-1]** and **[NE-2]**. In addition, each study tests **[Responsibility]**.

3.1 Study 1: Machine Case with Responsibility Attributions

Is it the case that people are unwilling to attribute responsibility to the norm-violating wire in the Machine Case? To test this, in my first study I solicited agreement ratings for both causal attributions and responsibility attributions. Each participant in Study 1 read Hitchcock and Knobe's original Machine Case vignette, then rated agreement with two pairs of attributions on separate pages—either a pair of causal attributions or a pair of responsibility attributions:

> **Cause:** The red wire caused the short circuit.
> The black wire caused the short circuit.
>
> **Responsible:** The red wire is responsible for the short circuit.
> The black wire is responsible for the short circuit.

The same 1-7 scale anchored at 1 with "strongly disagree," at 4 with "neither agree nor disagree," and at 7 with "strongly agree" was used for each study. The order of the two pages was varied and participants were not able to return to the previous page. The order of the two attributions on each page was randomized.

Participants for each study in this paper were recruited through advertising for a free personality test on Google Ads, with the personality test being administered after the target questions.[9] Responses were restricted to participants indicating that they are native English

[9] One notable benefit of using a "push strategy" like this one (i.e., recruiting participants who were not directly looking to participate in research) is that participants are more likely to be "experimentally naïve" and less likely to

speakers, 16 years of age or older, with at most minimal training in philosophy (completed an undergraduate major or more advanced studies). Responses for Study 1 were collected from 89 participants who met the restrictions.[10] Results are shown in Figure 1.

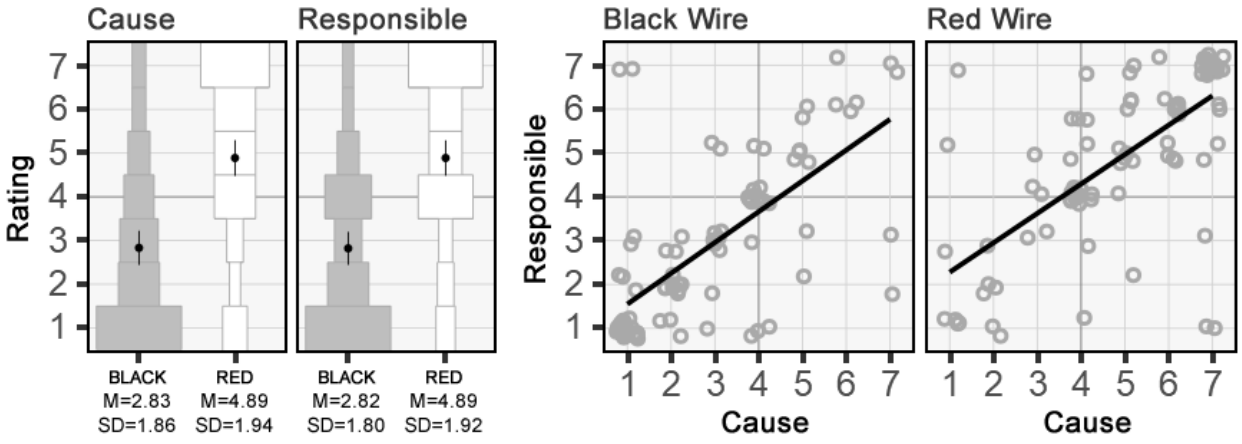

Figure 1: Results for Study 1. Plots on the left show relative percentage of participants selecting each response option, with means and 95% confidence intervals overlaid. Scatterplots on the right show points with jitter and regression lines calculated without jitter.

A three-way mixed ANOVA, with *page order* (Cause on first page, Responsible on first page) as a between-subjects factor and *term* (Cause, Responsible) and *wire* (Black, Red) as within-subjects factors, shows a main effect for *wire* [$F(1,87)=56.21$, $p<.001$, $\eta^2=.23$] and no other significant effects. In other words, the order of the pairs of attributions did not make a significant difference, nor did the type of attribution. Planned t-tests show that the original effect replicates [$t(88)=7.42$, $p<.001$, $d=.79$].[11] And, importantly, the same norm effect is also found for

be motivated to provide the responses that they think the experimenters are looking for (Haug 2018). Samples collected using the recruitment strategy employed here have been previously compared against samples collected with other methods in replication studies. And the present strategy has been consistently found to generate a diverse sample in terms of geography, socio-economic status, religiosity, political orientation, age, and education. Studies using this strategy have been previously reported in publications including, e.g., Livengood et al. (2010), Feltz and Cokely (2011), Sytsma and Machery (2012), Murray et al. (2013), Machery et al. (2015), Kim et al. (2016), Livengood and Rose (2016), Sytsma and Reuter (2017), Sytsma and Ozdemir (2019), Reuter and Sytsma (2020), Fischer et al. (forthcoming).

[10] 58.4% women, average age 30.8 years, ranging from 16 to 66.

[11] Two-tailed tests are used throughout, with Student's t-tests for one-sample or paired-sample comparisons and Welch's t-tests for independent-sample comparisons.

responsibility ratings [$t(88)=6.59$, $p<.001$, $d=.70$]. This speaks against prediction **[HK-1]**: the effect shown by Hitchcock and Knobe is *not* specific to causal attributions. As such, their finding for the Machine Case does not provide evidence of a norm effect in the absence of corresponding responsibility judgments and, hence, does not provide evidence favoring indirect accounts over direct accounts. In contrast, the results provide support for **[Responsibility]**: the ratings for Cause and Responsible are not statistically significantly distinguishable and they are very highly correlated (r=0.75).

3.2 Study 2: Further Attributions

In Study 2, I replicated the results of Study 1 and extended them to three further attributions:

Blame:	The red wire is to blame for the short circuit. The black wire is to blame for the short circuit.
Fault:	The short circuit is the red wire's fault. The short circuit is the black wire's fault.
Punish:	The red wire deserves to be punished for the short circuit. The black wire deserves to be punished for the short circuit.

Each participant read the original Machine Case vignette, then rated all ten attributions. The five pairs of attributions were either presented together on a single page in the order presented above or one pair was presented on a first page and the remaining four pairs presented on a second page in the same order. The vignette was repeated on each page and participants were not able to go back. To avoid confusion, the order of the attributions in each pair was the same as that shown above. Responses were collected from 242 participants meeting the restrictions.[12] The results are shown in Figure 2.

[12] 60.7% women (two non-binary), average age 31.5 years, ranging from 16-91.

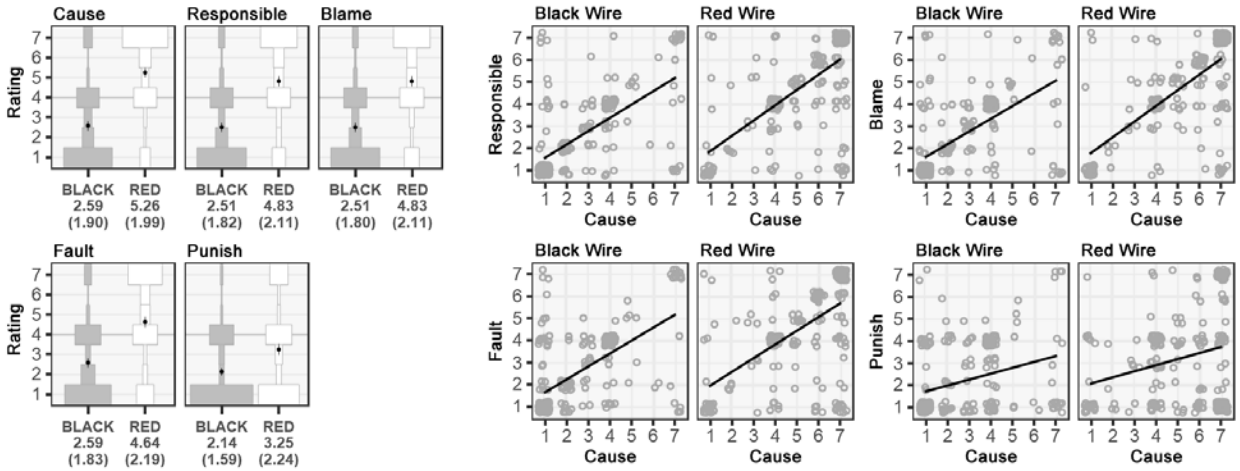

Figure 2: Results for Study 2. Plots on the left show relative percentage of participants selecting each response option, with means and 95% confidence intervals overlaid. Scatterplots on the right show points with jitter and regression lines calculated without jitter.

A three-way mixed ANOVA, with *order* (all pairs on first page, each individual pair alone on first page) as a between-subjects factor and *term* (Cause, Responsible, Blame, Fault, Punish) and *wire* (Black, Red) as within-subjects factors, shows main effects for *term* [$F(4,944)=51.84$, $p<.001$, $\eta^2=.035$] and *wire* [$F(1,236)=197.94$, $p<.001$, $\eta^2=.21$], as well as a significant interaction between the two [$F(4,944)=28.59$, $p<.001$, $\eta^2=.013$]. No other significant effects were found, including that there were no significant effects for *order*.

Planned t-tests reveal that Hitchcock and Knobe's original effect again replicates [$t(241)=14.28$, $p<.001$, $d=.92$], and a comparable norm effect is again found for responsibility attributions [$t(241)=12.70$, $p<.001$, $d=.82$]. These findings further indicate against prediction **[HK-1]**. More importantly, a significant norm effect is also found for blame attributions [$t(241)=12.38$, $p<.001$, $d=.80$], fault attributions [$t(241)=11.23$, $p<.001$, $d=.72$], and punishment attributions [$t(241)=7.54$, $p<.001$, $d=.48$].[13] These findings indicate against **[HK-2]**, further undermining Hitchcock and Knobe's argument against direct accounts. In addition, the same five

[13] Given that predictions are made about each pair of attributions, no corrections for multiple comparisons are applied for tests of the norm effects.

effects are found when restricting to just the responses when the pair of attributions was presented alone on the first page, ruling out that the presentation of all five pairs drove the effect: when considered alone there was a significant norm effect for Cause [$t(41)=4.48$, $p<.001$, $d=.69$], Responsible [$t(39)=4.04$, $p<.001$, $d=.64$], Blame [$t(39)=3.90$, $p<.001$, $d=.62$], Fault [$t(40)=4.82$, $p<.001$, $d=.75$], and Punish [$t(38)=2.09$, $p=.043$, $d=.34$].

Effect sizes are similar for the norm effects, except for Punish where the effect size is notably smaller. This is in line with prediction **[NE-2]**. More directly, in contrast to the other attributions, participants tended to deny that the red wire deserves to be punished, with the mean rating being significantly below the neutral point [$t(241)=5.19$, $p<.001$, $d=.33$]. Finally, in line with **[Responsibility]** there is once again an extremely strong correlation between causal attributions and responsibility attributions (r=.75). Similar correlations are found for blame (r=.75) and fault attributions (.69), while the relationship was notably weaker between causal attributions and punishment attributions (r=.38).

3.3 Study 3: Machine Case with Movement

The first two studies tested the original Machine Case vignette; but as discussed above, Samland and Waldmann note a potential confound: it might be that the (supposed) norm effect instead reflects that people assume that the position of the black wire is fixed, while the position of the red wire changes. To test this, I used a revised version of the Machine Case in which it was specified that both wires move around the machine:

> A machine is set up in such a way that it will short circuit if both the black wire and the red wire touch the battery at the same time. The machine will not short circuit if just one of these wires touches the battery. The machine is designed so that both wires move around inside the machine. The black wire is supposed to touch the battery at certain times as it moves around inside the machine. The red wire is never supposed to touch the battery as it moves around inside the machine.

> One day, the black wire and the red wire both come in contact with the battery at the exact same time. There is a short circuit.

After reading the vignette, participants were either asked to rate agreement with the Cause, Responsible, or Blame pairs from the previous study, now presenting them between-subjects (each participant rating one of the three pairs of attributions) with the order of the two attributions randomized. Responses were collected from 165 participants meeting the restrictions.[14] Results are shown in Figure 3.

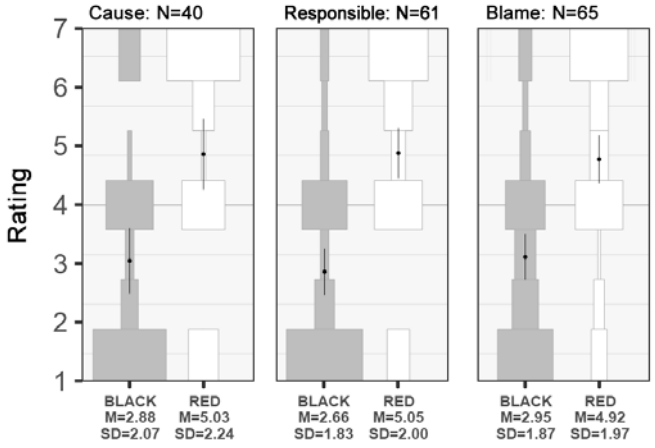

Figure 3: Results for Study 3. Plots show relative percentage of participants selecting each response option, with means and 95% confidence intervals overlaid.

A two-way mixed ANOVA with *term* (Cause, Responsible, Blame) as a between-subjects factor and *wire* (Black, Red) as a within-subjects factor, shows a main effect for *wire* [$F(1,163)=88.03$, $p<.001$, $\eta^2=.23$] and no other significant effects. Planned t-tests show that the norm effect occurs for Cause [$t(39)=4.88$, $p<.001$, $d=.77$], Responsible [$t(60)=6.12$, $p<.001$, $d=.78$], and Blame [$t(64)=5.27$, $p<.001$, $d=.65$], despite having revised the vignette to remove the confound noted by Samland and Waldmann. Thus, in addition to providing further evidence against **[HK-1]** and **[HK-2]**, the results indicate against **[SM]**. Once again, there is a close

[14] 63.6% women (one non-binary), average age 30.2 years, ranging from 16-70.

correspondence between Cause and Responsible (as well as Blame). In fact, responses for these attributions are not statistically significantly different, further supporting **[Responsibility].**

3.4 Study 4: Normative Evaluations

While the occurrence of the norm effect for plausibly normative terms like "responsible," "blame," "fault," and "punish" would seem to indicate that these terms are being applied in a normative sense, in Study 4 I test this more directly. Participants were given either the original Machine Case vignette or the revised vignette. This time, however, on the first page they were asked about one of two pairs of normative evaluations:

Wrong: The red wire did something wrong.
The black wire did something wrong.

Should Not: The red wire did something it should not have done.
The black wire did something it should not have done.

On a second page, participants rated the Cause, Responsible, and Blame pairs in this order. To avoid confusion, the order of the attributions in each pair was the same as that shown above. The vignette was repeated on the second page and participants were not able to go back. Responses were collected from 168 participants who met the restrictions.[15] Results are shown in Figure 4.

Starting with the first page, two-way mixed ANOVAs with *evaluation* (Wrong, Should Not) as a between-subjects factor and *wire* (Black, Red) as a within-subjects factor show a main effect for *wire* for both the original vignette [$F(1,82)=17.40$, $p<.001$, $\eta^2=.097$] and the revised vignette [$F(1,82)=30.27$, $p<.001$, $\eta^2=.15$] and no other significant effects. Planned t-tests show that the norm effect occurs for both normative evaluations for each vignette: ratings for the red wire are significantly greater than the black wire for Wrong [$t(41)=2.51$, $p=.016$, $d=.39$] and

[15] 78.6% women, average age 53.8 years, ranging from 16-92.

Should Not [$t(41)=3.50$, $p=.0011$, $d=.54$] for the original vignette, and for Wrong [$t(40)=3.11$, $p=.0034$, $d=.49$] and Should Not [$t(42)=4.59$, $p<.001$, $d=.70$] for the revised vignette.

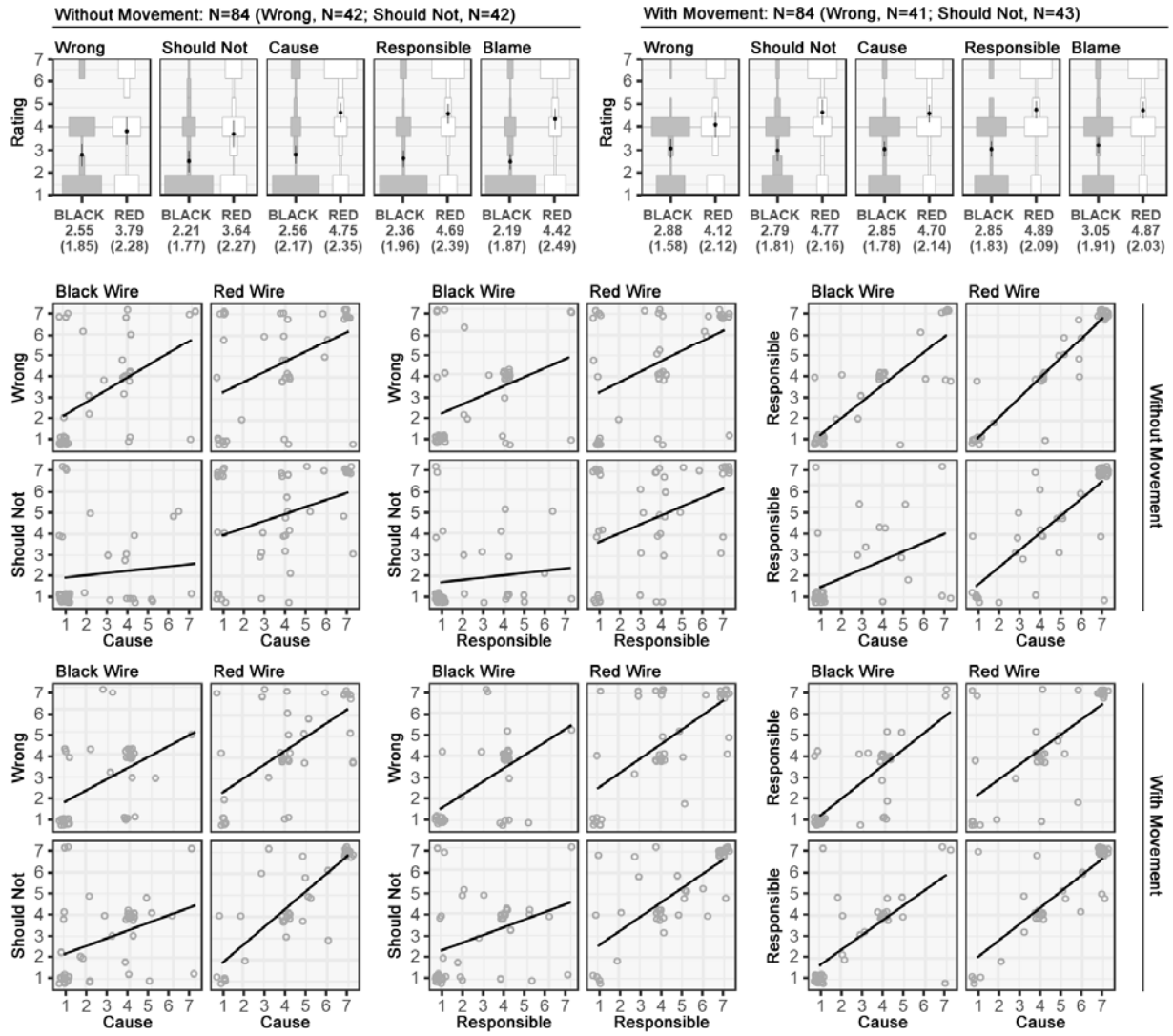

Figure 4: Results for Study 4. Plots above show relative percentage of participants selecting each response option, with means and 95% confidence intervals overlaid. Selected scatterplots below show points with jitter and regression lines calculated without jitter.

Turning to the attributions on the second page, two-way ANOVAs with *term* (Cause, Responsible, Blame) and *wire* (Black, Red) as within-subjects factors show a main effect for *wire* for both the original vignette [$F(1,83)=55.08$, $p<.001$, $\eta^2=.21$] and the revised vignette [$F(1,83)=47.58$, $p<.001$, $\eta^2=.19$], a main effect for *term* for the original vignette [$F(2,166)=3.98$,

p=.021, η^2=.003] although the effect size was negligible, and no other significant effects. Ratings for Cause are not significantly different from ratings for Responsible for either the black wire [t(83)=1.22, p=.23, d=.13] or the red wire [t(83)=.47, p=.64, d=.052] for the original vignette, however. Further, there is once again an extremely strong correlation between Cause and Responsible, both for the original vignette (r=0.85) and the revised vignette (r=0.80), providing further support for **[Responsibility]**. Planned t-tests reveal that that the norm effect again replicates for Cause for both the original vignette [t(83)=6.79, p<.001, d=.74] and the revised vignette [t(83)=6.22, p<.001, d=.68], and similarly for Responsible [t(83)=7.29, p<.001, d=.80; t(83)=6.69, p<.001, d=.73] and Blame [t(83)=7.02, p<.001, d=.77; t(83)=5.85, p<.001, d=.64]. As such, these findings further indicate against predictions **[HK-1]**, **[HK-2]**, and **[SM]**.

Comparing responses across the two pages, we see that the norm effect for Should Not for the revised vignette is of comparable size to the effects found for Cause, Responsible, and Blame. Further, there is a strong correlation between the normative evaluations on the first page and the attributions on the second page. This is especially pronounced for the revised vignette, where both Wrong and Should Not are very strongly correlated with Cause (r=0.63, r=0.73) and with Responsible (r=0.68, r=0.67). This provides initial support for **[Norm]**, suggesting that the norm effects are indeed *norm* effects.

To further test **[Norm]**, I performed a series of Bayesian within-subjects mediation analyses with 10k iterations (Vuorre and Bolger 2018). Analyses tested whether the normative evaluations mediate the norm effects. Since ratings for Cause and Responsible are not statistically significantly different for either vignette, these were combined for the analyses shown in Figure 5. For the original vignette, Wrong significantly mediates the norm effect, but Should Not is not a significant mediator. The results are even stronger for the revised vignette, with both normative evaluations mediating the norm effect. Most notably, Should Not fully

mediates the effect. In other words, the norm effect for Cause and Responsible is no longer significant when controlling for Should Not. These results indicate that the norm effect does indeed reflect participants' normative evaluations, providing strong support for **[Norm]**.

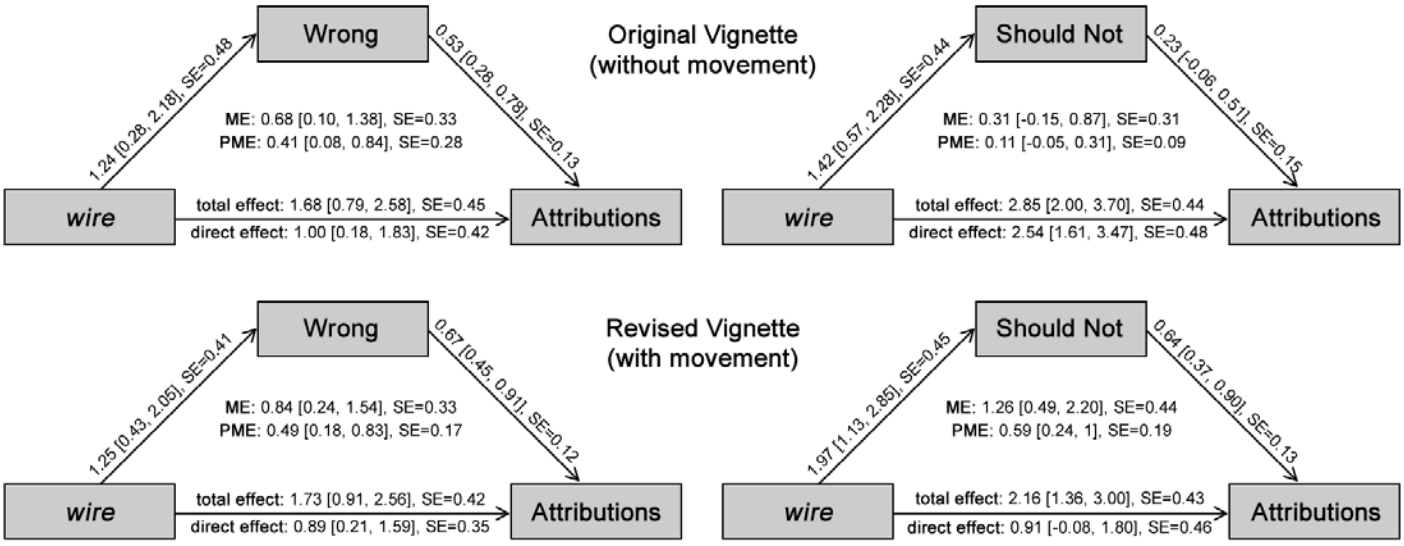

Figure 5: Results for mediation analyses for norm effect on averaged ratings for Cause and Responsible in Study 4, showing path diagrams with point estimates (posterior means) of the parameters, with standard errors, and associated 95% credible intervals, as well as the estimated direct effect, mediation effect (ME), and proportion of the effect mediated (PME).

3.5 Study 5: Normative versus Moral Evaluations

The results of Study 4 indicate that the effects in the Machine Case are in fact norm effects. But these results do not suggest between the agentive response and the normative response, since neither of the evaluations tested involve a clearly moral norm. To test between these explanations, in Study 5 I added a third evaluation that is clearly moral:

Moral: The red wire did something morally wrong.
The black wire did something morally wrong.

Each participant read the revised version of the Machine Case. On the first page, they rated one of the three normative evaluations (Should Not, Wrong, Moral) for *one* of the two wires. In other

words, unlike the previous studies this study used single evaluations rather than joint evaluations (see Sytsma ms for discussion). On the second page, participants then rated their agreement with either Cause or Responsible for the same wire rated on the first page. Again, the vignette was repeated on the second page and participants were not able to go back to the first page. Responses were collected from 729 participants who met the restrictions.[16] The results are shown in Figure 6.

Starting with the evaluations on the first page, a two-way ANOVA with *evaluation* (Should Not, Wrong, Moral) and *wire* (Black, Red) as between-subjects factors, shows main effects for *evaluation* [$F(2,723)=19.77$, $p<.001$, $\eta^2=.045$] and *wire* [$F(1,723)=102.38$, $p<.001$, $\eta^2=.12$] as well as a significant interaction [$F(2,723)=4.76$, $p=.0089$, $\eta^2=.011$]. Most importantly, planned t-tests show a significant norm effect for each of the three normative evaluations, although the effect size was much smaller for the distinctively moral evaluation, Moral [$t(248.96)=3.55$, $p<.001$, $d=.44$], compared to either Should Not [$t(237.74)=6.42$, $p<.001$, $d=.83$] or Wrong [$t(231.85)=7.70$, $p<.001$, $d=.99$]. Further, participants tended to deny that that "the red wire did something morally wrong," with the mean rating being significantly below the neutral point [$M=3.12$; $t(134)=4.65$, $p<.001$, $d=.40$], just as they tended to deny that "the red wire deserves to be punished" in Study 2, indicating that prediction **[NE-2]** holds.

[16] 77.4% female (four non-binary), average age 51.3 years, ranging from 16-90.

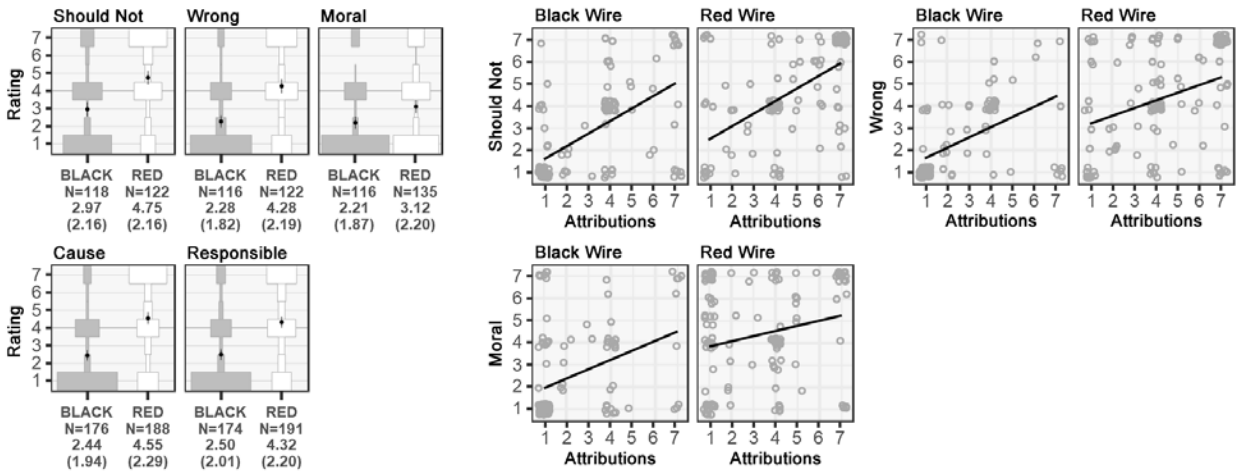

Figure 6: Results for Study 5. Plots on the left show relative percentage of participants selecting each response option, with means and 95% confidence intervals overlaid. Scatterplots on the right show points with jitter and regression lines calculated without jitter.

Turning to the attributions on the second page, a three-way ANOVA with *norm* (Should Not, Wrong, Moral), *wire* (Black, Red), and *term* (Cause, Responsible) as between-subjects factors, shows a main effect for *wire* [$F(1,717)=156.87$, $p<.001$, $\eta^2=.18$] and no other significant effects. In other words, the norm evaluated on the first page did not significantly affect attribution ratings, nor did the type of attribution rated, providing further support for **[Responsibility]**. Planned t-tests confirm that the norm effect is again found for both Cause [$t(358.32)=9.52$, $p<.001$, $d=.99$] and Responsible [$t(362.99)=8.29$, $p<.001$, $d=.87$], providing additional evidence against **[HK-1]** and **[SM]**.

Comparing responses across the two pages, we see that the norm effects for Should Not and Wrong are of comparable size to the effects found for Cause and Responsible, while the effect for Moral is notably smaller. And while ratings for each of the evaluations on the first page are correlated with ratings for the attributions on the second page, the correlation is notably stronger for Should Not (r=0.65) and Wrong (r=0.52) compared to Moral (r=0.35). The findings

for Should Not and Wrong provide additional support for **[Norm]**, while the contrast with Moral provides initial support for **[NE-1]**.

Finally, to further test **[Norm]** and **[NE-1]**, I performed bootstrap mediation analyses with 5k resamples (Preacher and Hayes 2008), looking at how each of the normative evaluations mediates the norm effect for the attributions. The results show that while each of the three evaluations significantly mediates the effect, Should Not mediates a very notably larger proportion of the effect (0.55) than does Moral (0.15), with Wrong in the middle (0.38), as seen in Figure 7. These findings provide further support for **[Norm]**, and most importantly provide support for **[NE-1]**: the normative but not distinctively moral judgments mediate the norm effect much more strongly than the moral judgment.

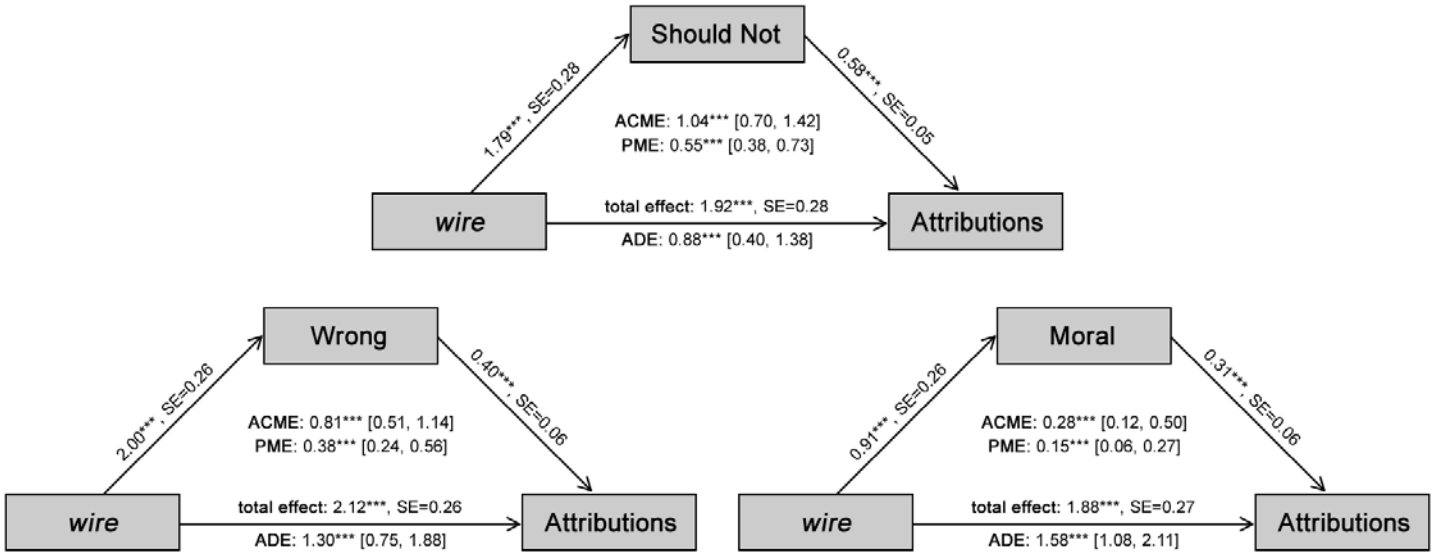

Figure 7: Mediation models with normative evaluations mediating the norm effect for Cause and Responsible. Unstandardized coefficients and standard errors are shown for each path. Average direct effect (ADE), average causal mediation effect (ACME), and average proportion of the effect mediated (PME) shown with associated 95% confidence intervals. Asterisks indicate significance, ***p<.001.

3.6 Summary

Across the five studies presented in this section, we find significant evidence that predictions **[HK-1]** and **[HK-2]** do not hold, with the norm effect occurring for a range of attributions, including responsibility and blame attributions, not just causal attributions. Further, we find evidence that the norm effects for the Machine Case are in fact *norm* effects: the results indicate against prediction **[SM]** derived from Samland and Waldmann's non-normative response, and the results bear out **[Norm]**, with the normative evaluations mediating the norm effects. In line with the predictions of the normative response, **[NE-1]** and **[NE-2]**, we find that the norm effects are primarily mediated by non-moral normative evaluations and that people tend to deny that the norm-violating wire deserves to be punished or did something morally wrong. Finally, each of the five studies supports **[Responsibility]**, with causal attributions and responsibility attributions showing a remarkable similarity across the studies.

Overall, these studies indicate against Hitchcock and Knobe's objection to direct views, and with it the agency assumption the objection is based on. Further, the evidence supports the normative response to Hitchcock and Knobe's objection, suggesting that people are willing to attribute responsibility and blame to non-agents because they tend to treat these attributions as being normative but not distinctively moral. And, within the direct views, the evidence fits most closely with the predictions of the responsibility view, favoring it not only over the counterfactual view, but the pragmatic and bias views.

4. Conclusion

Both philosophers and psychologists often assume that blame and responsibility only apply to certain agents. Sometimes this is nuanced by drawing a distinction between concepts of responsibility and blame, with one pair being taken to be purely descriptive and to apply to non-

agents (causal responsibility, causal blame) while the other pair is taken to be normative and to apply only to certain agents (moral responsibility, moral blame). This division, and the agent assumption more generally, is based not on evidence concerning ordinary concepts and their usage, however, but prior assumptions. In this paper, I've investigated one recent debate where this agent assumption has been wielded, focusing on discussions concerning the impact of injunctive norms on ordinary causal attributions.

Most prominently, Hitchcock and Knobe (2009) have argued against one family of accounts, and in favor of another, by presenting the results of a case where norms notably impact causal attributions but, they claim, blame attributions don't apply. This hinges on the agent assumption: they hold that blame attributions don't apply because the case involves artifacts. A further investigation of this case, however, provides clear evidence against this assumption: people are willing to ascribe blame and responsibility to the norm-violating artifact. Further, the results suggest that they do so because they are applying normative, but not distinctively moral, concepts. And the same holds for causal attributions. In fact, causal attributions and responsibility attributions were remarkably similar across the studies reported. Together the results suggest in favor of our responsibility view, and against competing accounts.

References

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

Eshleman, A. (2016). "Moral Responsibility." In E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*.

Feltz, A. and E. Cokely (2011). "Individual differences in theory-of-mind judgments: Order effects and side effects." *Philosophical Psychology*, 24(3): 343-355.

Fischer, E., P. Engelhardt, and J. Sytsma (forthcoming). "Inappropriate stereotypical inferences? An adversarial collaboration in experimental ordinary language philosophy." *Synthese*.

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

Haug, M. (2018). "Fast, Cheap, and Unethical? The Interplay of Morality and Methodology in Crowdsourced Survey Research." *Review of Philosophy and Psychology*, 9(2): 363-379.

Haywood, A. (2010). *Siberia: A Cultural History*. Oxford: Oxford University Press.

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

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.

Kim, H., N. Poth, K. Reuter, and J. Sytsma (2016). "Where is your pain? A Cross-cultural Comparison of the Concept of Pain in Americans and South Koreans." *Studia Philosophica Estonica*, 9(1): 136-169.

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.

Livengood, J. and D. Rose (2016). “Experimental Philosophy and Causal Attribution.” In J. Sytsma and W. Buckwalter (eds.), *A Companion to Experimental Philosophy*, Wiley Blackwell, 434-449.

Livengood, J., J. Sytsma, A. Feltz, R. Scheines, and E. Machery (2010). “Philosophical Temperament.” *Philosophical Psychology*, 23(3): 313-330.

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 (2020). “Actual causation and compositionality.” *Philosophy of Science*, 87(1): 43-69.

Machery, E., J. Sytsma, and M. Deutsch (2015). “Speaker’s Reference and Cross-cultural Semantics.” In A. Bianchi (ed.), *On Reference*, Oxford University Press, 62-76.

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

Murray, D., J. Sytsma, and J. Livengood (2013). “God Knows (But does God Believe?)” *Philosophical Studies*,166: 83-107.

Preacher, K. and A. Hayes (2008). “Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models.” *Behavior Research Methods*, 40: 879–891.

Reuter, K. and J. Sytsma (2020). “Unfelt Pain.” *Synthese*, 197: 1777-1801.

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

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

Schwenkler, J. and J. Sytsma (ms). “Reversing the Norm Effect on Causal Attributions.” http://philsci-archive.pitt.edu/18220/

Shaver, K. (1985). *The Attribution of Blame: Causality, Responsibility, and Blameworthiness*. New York: Springer.

Sytsma, J. (forthcoming). “Causation, Responsibility, and Typicality.” *Review of Philosophy and Psychology*.

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

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

Sytsma, J. and J. Livengood (ms). “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, J., and E. Machery (2012). “On the Relevance of Folk Intuitions: A Reply to Talbot.” *Consciousness and Cognition*, 21: 654-660.

Sytsma, J. and E. Ozdemir (2019). “No Problem: Evidence that the Concept of Phenomenal Consciousness is Not Widespread.” *Journal of Consciousness Studies*, 26(9-10): 241-256.

Sytsma, J. and K. Reuter (2017). “Experimental Philosophy of Pain.” *Journal of Indian Council of Philosophical Research*, 34(3): 611-628.

Tognazzini, N. and D. J. Coates (2016). “Blame.” In E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*.

Vuorre, M. and N. Bolger (2018). “Within-subject mediation analysis for experimental data in cognitive psychology and neuroscience.” *Behavior Research Methods*, 50: 2125–2143.