Please cite the final version as:

Güver, L., & Kneer, M. (2025). Causation, Norms, and Cognitive Bias. Cognition, 259, 106105. https://doi.org/10.1016/j.cognition.2025.106105

Causation, Norms, and Cognitive Bias

Abstract

Extant research has shown that ordinary causal judgments are sensitive to normative factors. For instance, agents who violate a norm are standardly deemed more causal than norm-conforming agents in identical situations. In this paper, we present novel findings that go against predictions made by several competing accounts that aim to explain this so-called "Norm Effect". By aid of a series of five preregistered experiments (N = 2'688), we show that participants deem agents who violate nonpertinent or silly norms – norms that do not relate to the outcome at hand, or for which there is little independent justification – as more causal. Furthermore, this curious effect cannot be explained by aid of potential mediators such as foreknowledge, desire or foreseeability of harm. The "Silly Norm Effect", we argue, spells trouble for several views of folk causality in the literature, and lends support to a Bias View, according to which Norm Effects are the result of blame-driven bias. We close with a discussion of the relevance of these findings for the just assessment of causation in the law.

Keywords: causation; norms; bias; blame; responsibility; foreseeability; negligence

1. Introduction

1.1 The impact of norms on perceived causation

A growing body of literature has revealed ordinary causal judgements to be susceptible to the violation of norms: when two agents jointly bring about an outcome, yet one does so in violation of a norm, the norm-violating agent is taken to be *the* cause of the outcome (Alicke, 1992, 2000; Hitchcock & Knobe, 2009; Knobe & Fraser, 2008; Kominsky et al., 2015; Samland & Waldmann, 2016; Samland et al., 2016; Icard et al., 2017; Cova et al. 2021; Henne et al., 2021; Olier, Willemsen & Kneer, 2025). We refer to this general phenomenon as the Norm Effect.[1] To illustrate, consider the following scenario: Mark is rollerblading on a footpath while Lauren is walking ahead of him. Suddenly, a cat jumps out of the brush, startling Lauren. She sidesteps into Mark's lane, who is unable to break in time. The two collide. When participants are confronted

[1] In the literature, the observation that agents, actions, or events are more frequently selected as causes when defying a norm is also sometimes referred to as "abnormal causal selection" (Gerstenberg & Icard, 2020; Henne et al., 2021) or "abnormal inflation" (Gill et al., 2022). Importantly, the Norm Effect has been shown to arise for a wide range of norms, including statistical norms, norms that prescribe behaviour, and norms of intended function (see e.g. Bear & Knobe, 2017; Gerstenberg & Icard, 2020; Kirfel & Lagnado, 2018; Knobe & Fraser, 2008; Livengood, Sytsma, & Rose, 2017; Morris et al., 2019; Sytsma, Livengood, & Rose, 2012). In our paper, we will use the locution "the Norm Effect" in a narrow sense, as pertaining to causal judgments concerning *agents* violating *prescriptive* norms. We will not dwell on statistical norms or norms of intended function.

with this version of the scenario and asked to label the cause of the accident, they overwhelmingly point to the cat.

Now consider a slight variation of the scenario in which everything is held fixed, except that there is now a norm in place that prohibits Mark from rollerblading on the footpath. Again, Lauren is startled by a cat, sidesteps into Mark's lane, and they collide. Who caused the accident? This time, participants overwhelmingly point to Mark: the violation of a salient norm has led to a drastic shift in their judgements (Güver & Kneer, 2023a). This difference in causal ascription across the no norm v. norm conditions is called the Norm Effect, and several accounts compete to explain it (for a review, see Willemsen & Kirfel, 2019; more generally, see Rose & Danks, 2012; Livengood & Rose, 2016; Henne, 2023; Bebb & Beebee, 2024). In what follows, we will give an overview of the four main types of account in the literature, before motivating the present experiments.

1.2 The Counterfactual and Pragmatic Views

According to proponents of the Counterfactual View, the Norm Effect results from the way in which counterfactuals figure in our causal judgements (Hitchcock & Knobe, 2009; Lagnado et al., 2013; Halpern & Hitchcock, 2015; Kominsky et al., 2015; Icard et al., 2017; Blanchard & Schaffer, 2017; Kominsky & Phillips, 2019; Henne et al., 2017, 2021; see also Phillips et al., 2015; Phillips & Knobe, 2018; Gerstenberg, 2024). Although the Counterfactual View comes in several flavours, the general idea goes as follows: when people are confronted with a norm violation, they are drawn to consider the counterfactual in which the agent adhered to the norm and evaluate the causal role of the norm-violating action in question. One example is the Necessity-Sufficiency Model proposed by Icard et al. (2017), which predicts that we judge the role of certain causal factors differently depending on whether the causal factor at hand was *normal* or *abnormal*. A causal factor is deemed abnormal, Icard and colleagues explain, when it violates a prescriptive or statistical norm, and normal when it does not violate either. According to the Necessity-Sufficiency Model, we judge normal causal factors on the basis of their *sufficiency* for the outcome, whereas we judge *abnormal* causal factors on the basis of their *necessity* for the outcome. Since prescriptive norm violations are standardly regarded as abnormal causal factors, the Necessity-Sufficiency Model predicts that participants assess the causal relations by looking to the *necessity* of the norm-deviant causal factor. In other words, participants consider the counterfactual in which the norm violation did not occur and probe whether the outcome would still have come about. The Norm Effect, then, is a result of participants' assessing the norm violation as causally necessary for the occurrence of the outcome. Proponents of this type of Counterfactual View would explain our introductory example as follows. If Mark had adhered to the prohibition and refrained from rollerblading on the footpath, the accident would not have come about. Since norm violating actions are judged on the basis of their *necessity* for the outcome, and Mark's action cannot be thought away without the outcome's also failing to obtain, Mark is deemed the cause of the ensuing accident.

Proponents of the Pragmatic View, by contrast, believe the expression "cause" to take different meanings depending on the conversational context in which it is employed (Samland & Waldmann, 2014, 2015, 2016; see also Samland et al., 2016). Sometimes people understand questions such as "Did Mark cause the accident?" descriptively, and answer by expressing a genuine causal judgement as to whether Mark *brought about* the accident or not. In other contexts, however, participants interpret the term "cause" normatively, and respond to questions such as "Did Mark cause the accident?" on the basis of their belief as to whether Mark ought to be held *accountable* for the accident. According to the Pragmatic View, then, the term "cause" is ambiguous: it is sometimes interpreted as referring to some descriptive, genuinely *causal* relation in the world, and at other times it is interpreted normatively to ascribe accountability to an agent. As regards our introductory example, proponents of the Pragmatic View hold that when participants are judging the norm-deviant Mark as more causal, they are not expressing a genuine causal judgement but want to express that Mark is morally responsible for the outcome. Thus understood, the Norm Effect is an artifact of language use, rather than of causal cognition.

1.3 The Responsibility and Bias Views

The Responsibility View holds that causal judgements are intimately tied to responsibility judgements, such that when ordinary people use locutions such as "Mark *caused* the accident", they take themselves to be saying something akin to "Mark is *responsible* for the accident" (Sytsma, 2019a, 2022; Sytsma et al., 2023; Sytsma, Livengood, & Rose, 2012). On the Responsibility View, the ordinary notion of "cause" does not denote an entirely descriptive relation between two entities in the world, but has a partially normative dimension (Sytsma, 2019b, 2021; Sytsma et al., 2012). The Norm Effect, in turn, is simply the upshot of the folk correctly applying this normative concept of causation (Sytsma, 2021). According to proponents of the Responsibility View, the norm violating agent has committed some type of transgression and is thus held more morally responsible – and subsequently, causally responsible – for the effect of her action than a norm-adhering agent in otherwise equal circumstances. Figure 1 illustrates the view.

Figure 1: A simple pathway model of the Responsibility View.

We turn, lastly, to the Bias View (of which Alicke's Culpable Control Model constitutes one kind, see Alicke 1992, 2000). The Bias View stands in stark opposition to the Responsibility View in that it deems the ordinary concept of causation to be descriptive, and regards the Norm Effect as a bias. As Alicke puts it, our "desire to praise or denigrate those whose actions we applaud or deride" gives rise to a performance error which distorts our causal judgements (Alicke et al., 2011). According to the Bias View, the Norm Effect constitutes one such distortion (Alicke et al. 2011;

Alicke & Rose, 2012; Rose, 2017; Rogers et al., 2019). To spell this out in a little more detail: Instead of *first* assessing an agent's potentially inculpating mental states as well as their causal contribution to a harmful outcome, and *thereafter* determining whether they are to be blamed, the process is frequently reversed: when assessing a norm-violating agent, we sometimes have an unconscious desire to blame them. To rationalize the desired blame ascription, we attribute the necessary constituents of blame, i.e. causal involvement and/or a guilty mind. Differently put, the process of moral judgment is conducted in reverse, and there is a considerable danger that we exaggerate the agent's causal contribution so as to justify the negative attitude we harbour towards them. The Bias View is schematised in Figure 2.

Figure 2: A simple pathway model of the Bias View.

1.4 Teasing apart the Responsibility and Bias Views

The Responsibility and Bias Views are, in many respects, very similar: they both posit that the Norm Effect is driven by normative judgements, be it those regarding responsibility or blame (which, after all, is just negative moral responsibility). How can they be distinguished? According to proponents of the Responsibility View, responsibility judgments are "broadly *moral* evaluations" (Sytsma, 2021). Unlike the Bias View, the Responsibility View requires us to distinguish "features that are *irrelevant* to appropriately assessing responsibility" (Sytsma, 2019b) from those that *legitimately* heighten the agent's responsibility for the outcome. Features that are irrelevant to the agent's responsibility – such as race, gender, sexual orientation, or general character (Alicke et al., 2011) – should, on the Responsibility View, not have an influence on causation, even if they inadequately influence *perceived* responsibility. The Bias View, on the other hand, does not draw a distinction between legitimate and illegitimate drivers of blame. It states that any feature apt to influence *perceived* blameworthiness – be it legitimate or illegitimate – can influence folk causal judgement. To tease apart the two views, one must thus explore whether factors irrelevant to moral responsibility proper influence causal judgement or not (Sytsma, 2019b).

An early example of this approach can be found in Alicke (1992). Alicke gave participants a vignette in which a speeding driver collides with another car, and manipulated the driver's motive for speeding. In one version, the driver was speeding in order to hide an anniversary gift for his parents; in the other, he was speeding to hide a vial of cocaine from his parents. Alicke's results seemed to suggest that persons with bad general character – a feature irrelevant to responsibility in the specific situation at hand (a road accident) – were indeed deemed more causal. However, as Sytsma (2019b) has suggested, the participants in Alicke's original study might have implicitly

drawn inferences from the agent's bad character to factors that *are* relevant to the agent's responsibility. In several replications, Sytsma illustrates that the difference between drivers speeding home – one to hide a present, the other to hide a vial of cocaine – is not only one of general character but also of perceived driving ability. Since a difference in driving ability *is* relevant to the assessment of agential responsibility when an accident occurs, Sytsma has argued that Alicke's original studies did not provide evidence against the Responsibility View.

Sytsma (2019b) has since constructed a more sophisticated version of the Responsibility View that accounts for the mediating role of several potentially inferred factors (Figure 3). Returning to our opening example, participants may, for instance, infer that Mark should have foreseen a crash (*foreseeability*) or did foresee it (*foresight*) when he violated a norm in skating on the footpath. In other scenarios, one might even go as far as inferring a *desire* to cause an accident. Just like ability or skill, the potentially inculpating mental state (*mens rea*) of the agent *is* relevant to the assessment of moral responsibility, and hence, on the Responsibility View, to the determination of causal responsibility.[2]

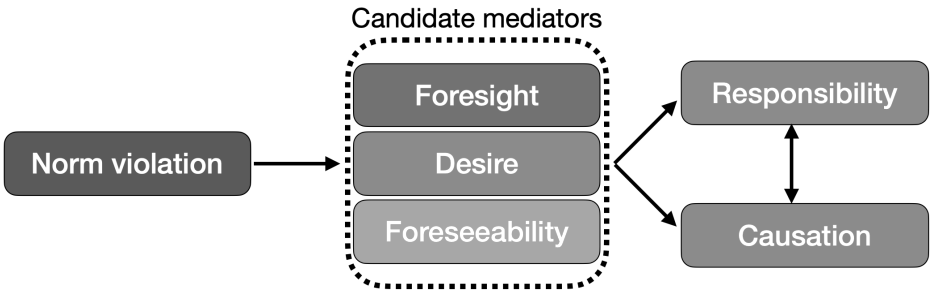

Figure 3: A more complex pathway model of the Responsibility View.

At this point it is helpful to distinguish two variations of the (complex) Responsibility View. According to a permissive version, *any* factor that impacts attributions of blame or responsibility may legitimate impact causation attributions (precisely because they impact responsibility). An account of this sort could be called the Anything-Goes View, since it is *too* permissive. It is too permissive because any factor that is able to sway *perceived* responsibility would constitute an adequate influence of causation: if attributions of blame or responsibility were, for instance,

[2] Kirfel and Philipps (2021, 2023) further highlight that the agent's mental state is relevant not only vis-à-vis the outcome, but also with respect to the norm violation itself. As Kirfel and Philipps have shown, agents who *unknowingly* violate norms are *not* judged more causal than their norm-adhering counterparts. Different accounts give different explanations of this phenomenon. Kirfel and Phillips, for instance, favour a counterfactual explanation: on their view, ignorance of a norm violation cancels the Norm Effect because people do not judge the norm-violating agent's behaviour as abnormal, and are consequently not drawn to imagine counterfactuals in which the agent does *not* violate said norm. In other words, they argue that where an agent is ignorant of their own norm-deviancy, we are less inclined to simulate alternatives to their norm-deviant behaviour. The Responsibility and Bias Views, too, are able to make sense of these findings. This is because agents who unknowingly violate norms are standardly neither responsible nor blameworthy for said norm violation, unless they *should have known* that they were violating a norm, *i.e.* acted negligently with respect to the norm. Since the agents in our experiments are all aware of their violating a norm, we can put this aspect of the debate aside. Nevertheless, Kirfel and Phillips' findings further emphasise the importance of controlling precisely for the mental states of the agent – a task which we take up in the following experiments.

influenced by gender in misogynistic ways, then on the Anything-Goes View, they would have a legitimate influence on causal attributions. This is, presumably, not how the Responsibility View is meant to be read. After all, its proponents have emphasised that certain factors are "peripheral" to the assessment of responsibility, such as, for example, the agent's gender or general character (Sytsma, 2019b). In contrast to the Anything-Goes View, then, the Responsibility View acknowledges that there are legitimate and illegitimate factors that may influence perceived responsibility, and that only the legitimate ones should impact causation. Hence, if adherence to salient norms – despite the fact that they have no clear connection to causation – are considered as a relevant factor for the assessment of moral responsibility, then their impact on causation is justified, too. However – and herein lies the difference to the Anything-Goes View – factors that should *not* influence perceived moral responsibility, such as race, gender, or general character, should *not* influence perceived causation either.

We are now in a position to see clearly what differentiates the Responsibility and Bias Views. The Responsibility View, unlike the Bias View, states that only *legitimate* drivers of responsibility should impact causal attributions. As Sytsma writes:

> 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 [i.e. *prima facie* irrelevant] 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. (2019b, p. 25)

The Responsibility and Bias Views, it appears, are in agreement that norm conformity is "peripheral" to causal attributions, *except* if it triggers *justifiable* inferences regarding mediators that correlate with moral responsibility. These mediators could be inculpating mental states (*e.g.* foresight, desire, negligence), the above-discussed abilities of the agent, or other factors that have a legitimate connection to moral responsibility.[3] Differently put, the Responsibility View seems to hold that a *direct* effect of norm violation on causal attributions is evidence in favour of the Bias View, whereas *indirect* effects via inculpating mental states and other "nonperipheral" factors support the Responsibility View. As Sytsma (2019b) has shown, once mental states such as foreknowledge and desire are explicitly controlled for, the (direct) effect of norm violations on causality attributions is marginal. These findings lend support to the Responsibility View, as they suggest that the effect of norms is not peripheral, as it exerts its influence not directly, but via the

[3] Naturally, the list of potential mediators is long and context-dependent. In the present paper, we are restricting ourselves to those mediators which proponents of the Responsibility View have themselves put forward (desire, foreknowledge), or which may plausibly reflect on the agent's moral responsibility (foreseeability).

mental states of the agent. Since the agent's mental states are relevant to her moral responsibility, on the Responsibility View, their influence on causation is legitimate.

1.5 The present experiments

In this paper, we would like to present two challenges to the Responsibility View, and shed further light on the Norm Effect more generally. As part of the first challenge, we will explore whether the violation of *nonpertinent* and *silly* norms gives rise to the Norm Effect. Nonpertinent norms, as we understand them, are norms which are sensibly in place, though whose violation does not stand in the appropriate relation to the ensuing harm. An example of this might be a norm that requires rollerbladers to wear helmets: although a sensible norm in its own right, its violation is of little importance in cases where the rollerblader collides with a pedestrian and the *pedestrian* gets injured. After all, helmets are meant to protect the person who is supposed to wear them. Silly norms, in turn, are norms that lack justification entirely: not only does their violation fail to relate to the ensuing harm, but there is little reason for them to be prescribed in the first place. An example would be a norm that allows rollerbladers to skate on the footpath, *unless* they happen to be wearing a grey shirt. This norm is generally silly, as it is not apparent what legitimate purpose it could serve (except, perhaps, in highly specific circumstances). Advocates of the Responsibility View are clear about the fact that neither nonpertinent nor silly norm violations should give rise to the Norm Effect, since it "is imperative for [the Responsibility View] that the norm-violating action is connected to the outcome" (Sytsma, 2019a, p. 14; see also Sytsma, 2022). Differently put, we take it that all parties to the debate agree that violations of nonpertinent or silly norm are "peripheral" factors, *i.e.* factors which clearly should not influence attributions of moral responsibility or causal contribution. If it turns out that they do, this would constitute evidence against the Responsibility View, and in favour of the Bias View.

Indeed, the challenge from nonpertinent and silly norms can also be directed at proponents of the Counterfactual View. Recall that the different Counterfactual Views explain the Norm Effect roughly as follows: when a norm violation occurs, people are inclined to think of counterfactuals in which the norm was adhered to, and notice that if the agent had *not* violated the norm, the outcome would not have obtained, or was less likely to obtain. The point of nonpertinent and silly norm violations is, of course, that they do *not* stand in any relation to the outcome: the violation of a nonpertinent or silly norm, in the absence of any additional explanatory factors, does *not* increase the likelihood of the outcome's occurrence. Thus, when participants simulate the counterfactual in which the norm violation did not occur, they should come to see that it is causally irrelevant to the outcome at hand. Hence, if the Norm Effect arose in nonpertinent and silly norm cases, this would put pressure on the Counterfactual View as well.

The most permissive account we have discussed, the Pragmatic View, by contrast, can accommodate the Norm Effect arising from nonpertinent or silly norms. The account predicts that

the expression "cause" is sometimes interpreted descriptively, and sometimes normatively. Whether normative interpretations, as in the nonpertinent and silly norm conditions, are *sensible* is – unlike for the Responsibility View – of no importance. Therefore, if attributed causality in the nonpertinent and silly norm conditions differ from no norm conditions, advocates of the Pragmatic View will simply argue that in the former, though likely not in the latter, the interpretation of "cause" carries normative charge.

The second challenge focuses on the effect of *pertinent* norms, *i.e.* norms which are appropriately connected to the outcome of interest. As proponents of the Responsibility View have argued – particularly, in response to the findings of Alicke (1992) – the violation of pertinent norms is only warranted if their effect on perceived causality is mediated by variables that appropriately influence moral responsibility (Sytsma, 2019b). The most evident variables of this sort are the agent's inculpating states of mind (*mentes reae*), such as intentionality or foresight, which have been explored by Sytsma. Here we try to show that there are cases where the influence of pertinent norm violations is *not* mediated by desire, foresight or foreseeability (see also Sytsma et al., 2012), and yet there remains, contrary to what the Responsibility View would predict, a residual Norm Effect.[4]

In sum, we aim to present a contingent objection to the Responsibility View (contingent on the mediators tested, and the precise account of the Responsibility View favoured), and a more direct challenge in the form of nonpertinent and silly norm violations. After all, nonpertinent or silly norm violations should impact neither moral responsibility nor causal attributions, be it directly or indirectly, because there are no reasonable features related to the downstream DVs that should be sensitive to an agent's adherence or violation of irrelevant or nonsense rules. And, as discussed, the latter challenge also applies to the Counterfactual View.

Here is how we will proceed. In Experiment 1, we show that nonpertinent and silly norm violations have a pronounced effect on participants' causal ascriptions, and that this effect cannot be explained by recourse to the proposed mediators of foreknowledge and desire. Experiment 2 builds on these findings and reveals that the effect of nonpertinent and silly norms disappears when the study takes a within-subjects design, *i.e.* when participants are presented the norm-adhering and norm-violating conditions side-by-side. This, we argue, shows that when participants are given a chance to reflect on the relevancy of the norm violation at hand, they recognise that nonpertinent and silly norm violations are not something that should factor into their causal judgement. Experiment 3 replicates the findings of Experiments 1 and 2 with a novel scenario. In Experiment 4, we pre-empt the criticism according to which foreknowledge and desire are the wrong mediators

[4] As a reviewer has helpfully pointed out, even if – to anticipate some of our later findings – it were true that the effect of norm violations is not mediated by mental states, it does not follow that the residual effect is therefore *direct*. Although the list of potential mediators is endless, the number of *sensible* mediators is rather small, and we hope to cover them in the experiments below.

and test instead whether participants deem the outcome foreseeable, *i.e.* whether they believe the agent acted negligently. While our *ex post* data suggests that participants do judge the accident as more foreseeable, our *ex ante* data reveals that participants fall prey to the hindsight bias, and once the hindsight bias is corrected for, the foreseeability of harm is unable to do the necessary explanatory work. We replicate these findings in Experiment 5, and close by considering their implications not only for the literature on the Norm Effect but also the law, where causal attributions play a central role in reaching a just verdict (see Knobe & Shapiro, 2021, Engelmann & Kirfel, 2024).

2. Experiment 1

In our first experiment, we explore both challenges to the Responsibility View empirically. We test whether the Norm Effect is mediated by inferred inculpating mental states (foreknowledge and desire), or whether they exert a (possibly direct) effect on perceived causation. We also explore whether the Norm Effect arises in cases where the agent violates nonpertinent or silly norms, *i.e.* norms that are unrelated to the outcome at issue or unrelated to *any* kind of potentially harmful outcome. The vignette, titled Festival, is based on a real criminal case.[5] All preregistrations, materials, data, and additional analyses are available under https://osf.io/24uvf/?view_only=ccd04f1940bd468eafd42757a2ea099b.

2.1 Participants

We recruited 305 participants on Amazon Mechanical Turk. Their IP address was restricted to the United States. As preregistered,[6] participants who failed an attention check, spent less than 10 seconds reading the vignette, failed a comprehension question, or were not native English speakers were excluded. 195 participants remained (exclusion rate: 36.1%; female: 45%; mean age: 40 years, SD = 12 years, range: 19–72 years).

2.2 Methods and materials

In the Festival scenario (full vignette in Appendix Section 4.1), Mark attends a music festival where Lauren is responsible for the special stage effects. During the concert, Lauren launches coloured powder over the dancing crowd which, unbeknownst to both her and the crowd, is combustible. The powder comes into contact with Mark's cigarette, ignites, and injures several festivalgoers.

The study took a between-subjects design and participants were randomly sorted into either the no norm, pertinent norm, nonpertinent norm, or silly norm condition. The no norm condition is silent

[5] See https://www.bbc.com/news/world-asia-33300970 (accessed 20 September 2023) and https://www.taipeitimes.com/News/front/archives/2016/04/27/2003644910 (accessed 20 September 2023).

[6] Available under https://aspredicted.org/FKE_PAO.

as to whether smoking is permitted on the festival grounds. In the pertinent norm condition, smoking is explicitly forbidden. In the nonpertinent norm condition, the festival organizers prohibited attendees to be topless. Nevertheless, Mark attends in his underwear only. In the silly norm condition, the festival – in an attempt to break a world record – had asked everyone to wear a green cap. Mark, who had initially agreed to do so, ultimately decides against it, and the festival fails to break the record.

Having read the vignette, participants were asked to report their agreement or disagreement with the following claims on 7-point Likert scales (labels in bold omitted):

> **Causation Mark:** "Mark caused the injuries." (1 = completely disagree; 7 = completely agree)
>
> **Causation Lauren:** "Lauren caused the injuries." (1 = completely disagree; 7 = completely agree)
>
> **Knowledge:** "Mark knew that the injuries would occur." (1 = completely disagree; 7 = completely agree)
>
> **Desire:** "Mark desired the injuries." (1 = completely disagree; 7 = completely agree)
>
> **Blame:** To what extent do you think Mark is blameworthy for the accident, if at all? (1 = not at all blameworthy; 7 = totally blameworthy)
>
> **Responsibility:** To what extent do you think Mark is *morally* responsible, if at all, for the accident? (1 = not at all morally responsible; 7 = totally morally responsible)
>
> **Punishment:** How much punishment, if any, does Mark deserve for the accident? (1 = no punishment at all; 7 = severe punishment)

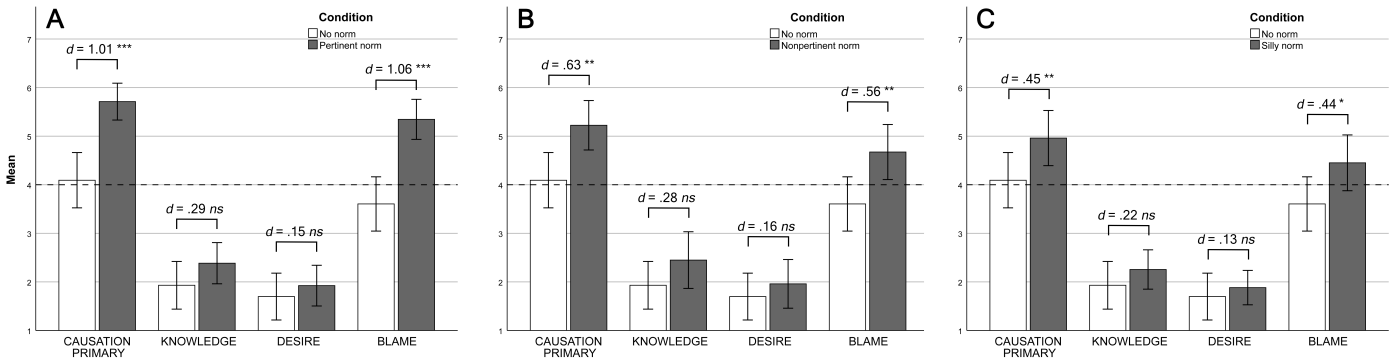

Figure 4: Comparison of means between the no norm and pertinent norm (A), nonpertinent norm (B), and silly norm (C) conditions. Effect sizes are given in terms of Cohen's d, * indicates $p < .05$, ** indicates $p < .01$, *** indicates $p < .001$. Error bars denote 95% confidence intervals.

2.3 Results

ANOVAs. One-way ANOVAs revealed a significant influence of norm status (no norm, pertinent, nonpertinent, and silly norms) on Mark's perceived causal contribution, blame, responsibility, and deserved punishment (all *p*s < .011). The effect sizes for Mark's causal contribution and blameworthiness were large (η_p^2s > .156) and the effect sizes for responsibility and punishment moderate (responsibility: $\eta_p^2 = .126$; punishment: $\eta_p^2 = .115$). Importantly, the effect of norm status on knowledge and desire proved nonsignificant (*p*s > .234). We further ran planned comparisons for all dependent variables, contrasting each of the three norm types with the no norm condition.

No norm v. pertinent norm. A comparison of the no norm and pertinent norm conditions revealed that participants in the pertinent norm condition judged Mark significantly and pronouncedly more causal, blameworthy, responsible, and deserving of punishment (*p*s < .001, *d*s > 1.00, large effects). There was no statistically significant difference across conditions for the knowledge and desire variables (*p*s > .159; see Figure 4A and, for this and the following contrasts, Appendix Section 3.1).

No norm v. nonpertinent norm. A comparison of the no norm and nonpertinent norm conditions yielded surprising results: in the nonpertinent norm condition – where all Mark did was violate the dress policy – participants rated him more causal, blameworthy, and responsible for the accident, as well as more deserving of punishment (all *p*s < .009, all *d*s > .55). The difference in knowledge and desire ratings did not reach statistical significance (*p*s > .173; see Figure 4B).

No norm v. silly norm. A contrast of the no norm and silly norm conditions revealed the previous pattern to persist: participants judged Mark significantly and considerably more causal, blameworthy, responsible, and deserving of punishment (all *p*s < .039, all *d*s > 0.43, moderate

effects). There was no statistically significant impact of norm type on knowledge and desire judgements (*p*s > .302; see Figure 4C).

2.4 Discussion

Our experiment replicated previous findings concerning the Norm Effect: Mark was judged more causal in the condition where he violated a pertinent norm vis-à-vis the condition where he did not violate any norm. However, knowledge and desire ratings were unaffected by norm violation, and significantly below the midpoint (and hence unlikely inferred factors). If one were to hold, as proponents of the Responsibility View seem to, that the Norm Effect can only be accommodated if there is a further, reasonable mediating factor – like *mens rea* – triggered by the difference in norms, then our findings count as tentative counterevidence to this account (tentative because there might be other, untested factors).

Our experiment also replicated preliminary findings according to which nonpertinent and silly norms exert an effect on blame and causation (Güver & Kneer, 2023a, 2023b). Recall that the nonpertinent and silly norm conditions were explicitly designed so that violating them would not make the agent more morally responsible for the outcome, given the clear lack of connection between the norm violation (*e.g.* failing to adhere to the dress code) and the harm that ensued (the injury of some festivalgoers). Our results, however, revealed that participants *do* deem the nonpertinent and silly norm violating agent as more causal and more blameworthy. These findings directly challenge the Responsibility View, and put some pressure on the Counterfactual View as well. They challenge the Responsibility View because, as its proponents have stressed, in the absence of a connection between action and outcome, an agent should not be held responsible (Sytsma, 2019a). They also put pressure on the Counterfactual View: in the relevant counterfactual in which the norm violation does *not* occur, the outcome would remain unchanged. Hence, the asymmetry in attribution cannot be explained by aid of an asymmetry across the actual and the counterfactual cases. Contrary to what the Necessity-Sufficiency Model would predict, then, the Norm Effect cannot – at least in this case – be driven by participants' being drawn to consider the necessity of the norm-violating action in question, because the norm violation was, in fact, entirely peripheral to the occurrence of the outcome. Overall, neither the Responsibility View nor the Counterfactual View would predict the Norm Effect to arise with nonpertinent or silly norm violations. Yet this is exactly what we find: Participants judged Mark in the nonpertinent and silly norm conditions as pronouncedly more causal, blameworthy, responsible, and deserving of punishment than in the norm-adhering condition – and this, despite there being no significant difference in desire or knowledge ascriptions.[7]

[7] For the role of knowledge – or lack thereof – in ascriptions of causation and responsibility, see *e.g.* Samland et al. (2016); Kirfel & Lagnado (2021c); Engelmann (2022); Kirfel et al. (2023).

The findings, we think, are most naturally interpreted as consistent with the Bias View: Participants view the norm-violating agent in a negative light, due to the disregard he has demonstrated for the societal norms in place – irrespective of how irrelevant or silly those norms are. This triggers a desire to blame the agent and, in an attempt to justify such blame, participants exaggerate his causal contribution.

There is, however, an alternative explanation of the findings. Perhaps the folk *do* view nonpertinent and silly norm violations as relevant to the agent's moral responsibility. In other words, there could be a difference between what moral philosophy and folk ethics deem normatively appropriate. If this were the case, then proponents of *some* version of the Responsibility View – what we have labelled the Anything-Goes View above – could argue that the difference in perceived causality is, after all, justified. This is because on the Anything-Goes View, *any* factor that influences responsibility judgements – be it legitimately or illegitimately – may impact causal ascriptions.

In Experiment 2, we seek to explore this question in more detail. In particular, we want to explore whether participants, when given the opportunity to further reflect on the norm violation at hand, continue to regard the nonpertinent or silly norm-violating agent as more responsible and maintain that this merits a difference in causation judgements. In order to do so, we ran Experiment 1 in a within-subjects design, confronting participants with two conditions side-by-side: each participant was given the no norm condition in addition to one of the norm violation conditions (pertinent norm, nonpertinent norm, silly norm), so as to, presumably, incite more reflection. The idea behind employing a within-subjects design is to make salient the single difference across the conditions (the type of norm violated) in order to see whether nonpertinent or silly norm violations do, according to what ordinary people themselves think, merit a difference in attributed moral responsibility (or blame) and causation. If the Norm Effect on nonpertinent and silly norms is a bias, then we would expect it to disappear when participants are (implicitly) invited to reflect more deeply on the status of the norm. This expectation is in line with previous work by Pinillos et al. (2011), who have found that contrastive designs of this sort allow participants to make more informed judgements (in their words, puts participants in a "better epistemic position"). Furthermore, within-subjects designs have been fruitfully employed as debiasing tools in moral psychology more broadly: for instance, in studies investigating moral luck, the sizeable between-subjects effect of outcome (neutral v. bad) on blame largely disappears in within-subjects designs, suggesting that it *does* constitute a bias (Kneer & Machery, 2019; Frisch et al. 2021; Kneer & Skoczen, 2023; see also Baron, 2008; Hsee, 1996).

3. Experiment 2

Experiment 2 explores whether a more permissive version of the Responsibility View – the Anything-Goes View – can explain the results reported in Experiment 1. While the Responsibility View requires ascriptions of responsibility to be *justified*, on the Anything-Goes View, any

perceived driver of responsibility may impact causation judgements. The aim of Experiment 2 is to see whether participants, when given the opportunity to further reflect on the norm violation by means of a contrastive design, maintain that the nonpertinent or silly norm-violating agent is more responsible for the outcome, or whether they come to realise that their knee-jerk reaction to blame the agent is unfounded. If, as we hypothesise, the Norm Effect of nonpertinent and silly norms were to disappear, this would provide further evidence for the Bias View.

3.1 Participants

358 participants were recruited online via Amazon Mechanical Turk. Their IP address was restricted to the United States. As preregistered,[8] participants who failed an attention check, spent less than 20 seconds reading the vignette, or were not native English speakers were excluded. 287 participants remained (exclusion rate: 19.8%; female: 49.5%; mean age: 44 years, SD = 14 years, range: 21–84 years).

3.2 Methods and materials

The study, building on the Festival vignette introduced above, took a mixed-factorial design (within-subjects factor – norm status: no norm v. norm; between-subjects factor – norm type: pertinent v. nonpertinent v. silly). It was identical to Experiment 1 in all respects, except that participants were presented with *two* conditions on the same page (no norm on the one hand, and pertinent, nonpertinent, and silly norm on the other) and were subsequently asked to judge all measures with respect to *both* conditions. We distinguished the two conditions by referring to the primary and secondary agents in the no norm condition as "Mark" and "Lauren", and the primary and secondary agents in the contrastive conditions as "John" and "Mary". Their presentation order was fixed, such that participants always read the no norm condition, followed by one of the norm-violating conditions (on the same page).

Participants were asked the same questions as in Experiment 1. That is, they were asked to rate the extent to which the primary (Mark and John) and secondary (Lauren and Mary) agents causally contributed to injuring the festivalgoers, whether the primary agents had any foreknowledge or desire as to the harm, and, finally, how blameworthy, morally responsible, and deserving of punishment the primary agents were. As in Experiment 1, all responses were recorded on 7-point Likert scales.

3.3 Results

[8] Available under https://aspredicted.org/3T4_JWR.

3.3.1 General Results

ANOVAs. Repeated-measures ANOVAs revealed a significant effect of norm status on the causal contribution of the primary agent ($p < .001$, $\eta_p^2 = .070$) and the moral variables of blame, responsibility, and punishment (ps $< .001$, η_p^2s $< .110$). The effect on knowledge and desire, however, was very small (η_p^2s $< .020$) and reached significance only for knowledge ($p = .020$). The effect of norm status on desire was nonsignificant ($p = .062$).

We ran planned contrasts for a more detailed breakdown of the impact of norm type on the dependent variables. Table 1 contrasts a summary of the key findings with the between-subjects findings from Experiment 1 (for full tables and analyses, see Appendix Section 3.2).

		Between-subjects					Within-subjects				
Contrast	*Variable*	*df*	*t*	*p*	Cohen's *d*	95% CI	*df*	*t*	*p*	Cohen's *d*	95% CI
NN v. PN	Causation Primary	76	–4.77	<.001	1.01	[–1.44;–.58]	91	–6.72	<.001	0.70	[–.93;–.47]
	Knowledge	93	–1.42	0.160	0.29	[–.70;.12]	91	–2.68	0.009	0.28	[–.49;–.07]
	Desire	93	–.71	0.478	0.15	[–.55;.26]	91	–1.80	0.075	0.19	[–.39;.02]
	Blame	93	–5.15	<.001	1.06	[–1.49;–.63]	91	–8.42	<.001	0.88	[–1.12;–.64]
NN v. NP	Causation Primary	90	–3.00	0.004	0.63	[–1.04;–.21]	93	–.84	0.403	0.09	[–.29;.12]
	Knowledge	89	–1.37	0.174	0.28	[–.69;.13]	93	–.75	0.455	0.08	[–.28;.13]
	Desire	90	–.75	0.454	0.16	[–.57;.25]	93	–.80	0.428	0.08	[–.28;.12]
	Blame	90	–2.69	0.008	0.56	[–.98;–.14]	93	–.53	0.597	0.06	[–.26;.15]
NN v. SN	Causation Primary	92	–2.16	0.034	0.45	[–.86;–.03]	100	0.42	0.675	0.04	[–.15;.24]
	Knowledge	92	–1.04	0.303	0.22	[–.62;.19]	100	–.46	0.650	0.05	[–.24;.15]
	Desire	92	–.63	0.528	0.13	[–.54;.28]	100	–.67	0.503	0.07	[–.26;.13]
	Blame	92	–2.11	0.038	0.44	[–.85;–.02]	100	–.18	0.855	0.02	[–.21;.18]

Table 1: Comparison of effect sizes for the no norm v. pertinent norm (NN v. PN), nonpertinent norm (NN v. NP), and silly norm (NN v. SN) conditions across ascriptions of causation, mental states, and blame. 95% confidence intervals for the reported *d*-values.

3.3.2 Planned comparisons

No norm v. pertinent norm. Participants judged the primary agent in the pertinent norm condition, John, as more causal than his no norm counterpart, Mark ($p < .001$, $d = .70$) (see Figure 5a). They also judged John to be significantly more blameworthy, morally responsible, and deserving of punishment than Mark (ps $< .001$, ds $> .77$). The effects on knowledge and desire were small (ds $< .29$) and reached significance only for knowledge ($p = .009$). The effect sizes for the core variables causality and blame are significantly smaller in the within-subjects design than the between-subjects design (see Appendix Section 3.2). Only about one in three participants rated the causal contribution of Mark, as well as blame and responsibility identically across scenarios. This suggests that a significant majority, when viewing the two scenarios side-by-side, considered the pertinent norm violation a *legitimate* influence on causation, blame and responsibility.

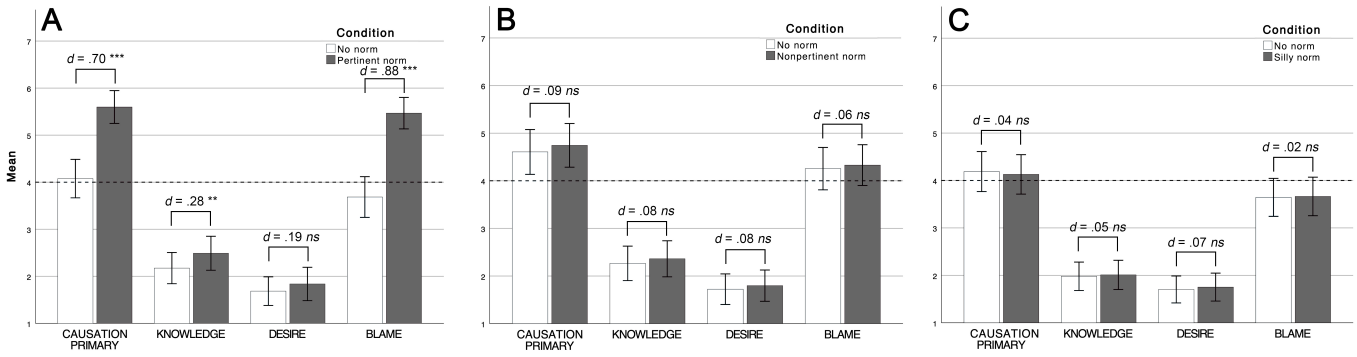

Figure 5: Comparison of means between the no norm and pertinent norm (A), nonpertinent norm (B), and silly norm (C) conditions. Effect sizes are given in terms of Cohen's d, * indicates $p < .05$, ** indicates $p < .01$, *** indicates $p < .001$. Error bars denote 95% confidence intervals.

No norm v. nonpertinent norm. In comparing the no norm and nonpertinent norm conditions (see Figure 5b), we did not find any statistically significant differences in participants' assessments of the dependent variables (all *p*s > .173), with very small effect sizes throughout (all *d*s < .15). The effect sizes for the core variables causality and blame are significantly smaller in the within-subjects design than the between-subjects design (see Appendix Section 3.2). By contrast to the pertinent norm comparison, more than 70% of the participants rated Mark's causal contribution, deserved blame and his moral responsibility identically across conditions. This suggests that most people considered the nonpertinent norm as *irrelevant* for the assessment of these dependent variables.

No norm v. silly norm. A comparison of the no norm and silly norm conditions, too, did not yield any statistically significant differences for the dependent variables (all *p*s > .123), with tiny effect sizes throughout (all *d*s < .08, except for the secondary agent's causal contribution at $d = .15$), see Figure 5c. The effect sizes for the core variables causality and blame are significantly smaller in the within-subjects design than the between-subjects design (see Appendix Section 3.2). More than 80% of the participants rated Mark's causal contribution, deserved blame and his moral responsibility identically across conditions, again suggesting that they considered it as irrelevant for the latter's assessment.

3.4 Discussion

With respect to the nonpertinent and silly norm conditions, the results paint a clear picture: Whereas in the between-subjects comparisons (Experiment 1) there were significant and medium-to-large effects for causation and the moral variables, the effects vanished in the within-subjects comparisons (Experiment 2), see Table 1 and Figure 5. Upon reflection, more than two thirds of the participants did not judge nonpertinent and silly norm-violating agents differently from norm-

conforming ones (see Appendix Section 3.2, Table 10). This suggests that participants realised both that the nonpertinent and silly norm violations were irrelevant to the outcome at hand, and also that they were irrelevant to the assessment of the dependent variables. After all, the data suggests that according to lay participants themselves, violations of nonpertinent and silly norms – at least if these are evident as the only difference across cases – should *not* impact causation judgments, and hence the effects that arise in between-subjects designs must be considered a bias. These findings for causation track the results of between- and within-subjects contrasts on *mens rea* attributions reported by Kneer & Machery (2019) and Kneer & Skoczen (2023). Here too, pronounced between-subjects effects of outcome on negligence and blame disappear once people see both cases side-by-side, suggesting an outcome bias.

The situation is more complex in the case of pertinent norms and allows interesting insights into the Norm Effect as discussed in the literature more generally. In the pertinent norm condition, too, we find a reduction in effect size across all variables. The effect on causation, for instance, drops from large (d = 1.01) in the between-subjects design to medium-sized (d = .70) in the within-subjects design, a statistically significant reduction (see Appendix Section 3.2, Table 16). Additionally, one third of the participants gave identical ratings to the causation and blame questions across the no norm and pertinent norm conditions (Appendix Section 3.2, Table 9). Inciting reflective judgement via a within-subjects design thus significantly reduces the Norm Effect.

The significant reduction of the Norm Effect for pertinent norms in the within-subjects design, and the fact that it entirely vanishes for about one third of our participants, suggests, that its size and extent is at least *partially* driven by bias: When facing the two conditions (no norm v. pertinent norm) side-by-side, many participants do not consider norm violations of much, or any, importance to the assessment of causality. Nevertheless, our findings suggest that a residual – and considerable – Norm Effect persists. When it comes to pertinent norms, participants judge the norm-violating agent as more causal and blameworthy, even in direct comparison to a norm-adhering agent. This suggests that the *residual* Norm Effect in the context of pertinent norm violations might in fact not result from biased reasoning. The different accounts under discussion can all accommodate this finding. Proponents of the Responsibility View may, for instance, would treat it as evidence for a partially normative concept of causation, and point to the strong correlation between perceived causation and moral responsibility – both in the no norm condition (r = .73), and the pertinent norm condition (r = .55) – in order to bolster their case (see Appendix Section 3.2). Proponents of the Counterfactual View could argue that the residual Norm Effect arises because participants are drawn to consider relevant alternatives (relevant, because the norm violation in question is *pertinent* to the outcome) and recalibrate their assessment on the basis of the necessity and sufficiency of the norm-deviant behaviour. Advocates of the Pragmatic View, who have the most legroom, would simply argue that in the no norm v. pertinent norm contrast, some participants favour the same (descriptive) interpretation of "cause" across conditions. About two thirds, by

contrast, do not – which explains the asymmetry in causality attributions in the within-subjects design.

4. Experiment 3

In order to explore the external validity of the results so far, we reran Experiments 1 and 2 with a different scenario (Trash Bag). Our aim was to see whether the following findings would replicate: (i) the curious – and pronounced – effects for nonpertinent and silly norm infractions on causality and blame attributions in between-subjects designs (Experiment 1), (ii) their independence from *mens rea* mediators invoked by proponents of the Responsibility View (Experiment 1), and (iii) the substantial decrease in effect size of all DVs in within-subjects designs and hence their interpretation as biases (Experiment 2). Since the experiments are direct replications of Experiments 1 and 2, we will be relatively concise.

4.1 Participants

We recruited participants for the two sub-experiments separately on Amazon Mechanical Turk, restricting the IP address to the United States. For the between-subjects design, 283 participants were recruited. As preregistered,[9] we excluded participants who failed an attention check, spent less than 15 seconds reading the vignette, gave a wrong answer to the comprehension question, or were not native English speakers. 212 participants remained (exclusion rate: 25.1%; female: 47%; mean age: 43 years, SD = 13 years, range: 22–75 years). For the within-subjects design, 396 participants were recruited. In line with our preregistration criteria,[10] we excluded participants who failed an attention check, spent less than 25 seconds reading the vignette (which was longer than in the between-subjects design), or were not native English speakers. 354 participants remained (exclusion rate: 10.6%; female: 50%; mean age: 44 years, SD = 13 years, range: 20–76 years).

4.2 Methods and materials

Participants received a short vignette based on a German Imperial Court of Justice (*Reichsgericht*) case.[11] In the scenario Mark places several trash bags outside his apartment building. Nearby, construction workers are cutting concrete with a buzz saw. The sparks light the trash bags ablaze, which results in the apartment building catching fire. Several tenants are injured. (The complete Trash Bag scenario can be found in the Appendix, Section 4.2).

[9] Available under https://aspredicted.org/DIZ_UZQ.
[10] Available under https://aspredicted.org/KJQ_FNW.
[11] RGSt 61, 318.

In the no norm condition, Mark was free to store his trash bags at the building's entrance. In the pertinent norm condition, city regulations prohibited the storing of objects near building entrances. In the nonpertinent norm condition, although the city required its citizens to use blue trash bags, Mark continued to use grey ones, which were identical in all properties but colour. In the silly norm condition, due to the abundance of sailors living in Mark's apartment building, all tenants were required to tie their trash bags with a special sailor's knot, which Mark did not do.

In the *between-subjects* design, participants were randomly assigned to one of the four norm conditions (no norm, pertinent norm, nonpertinent norm, silly norm). In the *within-subjects* design, participants were randomly assigned to pairs of scenarios contrasting the no norm condition with one of the three norm conditions on the same screen (as in Experiment 2, their presentation order was fixed). They were, as in the previous experiments, asked to rate the causal contributions of the primary agent, Mark, and the secondary agents (the workers). They were further asked to judge Mark's foreknowledge of, and desire to, bring the outcome about, as well as the moral variables of blame, responsibility, and punishment. As in Experiments 1 and 2, all items were presented on 7-point Likert scales.

4.3 Results

4.3.1 General results

Between-subjects design. One-way ANOVAs investigating the influence of norm type (no norm v. pertinent norm v. nonpertinent norm v. silly norm) revealed a significant main effect on the causal contribution of Mark and the moral variables of blame, responsibility, and punishment (all ps < .001), with large effect sizes throughout (η_p^2s > .242). The effect of norm type on knowledge and desire proved nonsignificant, though knowledge was close (p = .058).

Within-subjects design. We ran repeated-measures ANOVAs to explore the influence of the three types of norms (pertinent v. nonpertinent v. silly) on the dependent variables. Aggregating across the three norm type conditions, we found participants' causal ascriptions to differ significantly with respect to the norm-violating agent (p < .001, η_p^2 = .163, a large effect). The difference in mental state ascriptions was small (η_p^2s < .048) and reached significance only for knowledge (p < .001). The moral variables, on the other hand, all differed significantly and pronouncedly across conditions (all ps < .001, all η_p^2s > .132).

4.3.2 Planned comparisons

For each design, we ran planned comparisons for a more detailed breakdown of the impact of norm type on the dependent variables, see Table 2 (complete tables in Appendix Section 3.3).

No norm v. pertinent norm. In the between-subjects design, the pertinent norm significantly increased attributions of causality, blame, responsibility and deserved punishment (all *p*s < .001, all *d*s > 1.73). In the within-subjects design, we also found significant and pronounced differences for these variables (all *p*s <. 001). As a meta-analysis across designs shows, however, the effect sizes – though they remained substantial (all *ds* > .78) – were reduced significantly to about half for causation and the moral variables (see Appendix Section 3.3.2 for full results). There was no significant effect on desire in either design (*p*s > .057, *ds* < .18), whereas there was a small-to-medium sized effect on knowledge in both (*p*s < .016, *ds* < .49). In the within-subjects design, the proportion of identical responses for causation was 42%, for blame 31%.

No norm v. nonpertinent norm. In the between-subjects design, the nonpertinent norm significantly increased attributions of causality, blame, responsibility and deserved punishment (all *p*s < .001, all *d*s > 1.07, large effects). In the within-subjects design, we also found a significant, yet significantly smaller effect for these variables (all *p*s <. 001, *ds* > .34); the effect size was reduced to at most half of the between-subjects effect (see Appendix Section 3.3.2 for a comparison of effect sizes). We could find no significant effect on desire or knowledge in either design (*ps* >. 477). In the within-subjects design, the proportion of identical responses for causation was 74%, for blame 78%.

No norm v. silly norm. In the between-subjects design, the silly norm significantly increased attributions of causality, blame, responsibility and deserved punishment (all *p*s < .001, all *d*s > .88, large effects). In the within-subjects design, none of the effects reached significance (all *ps* > .055, all *ds* < .19), and they were significantly smaller than in the between-subjects design (see Appendix Section 3.3.2). We could find no significant effect on desire or knowledge in either design (*ps* > .091). In the within-subjects design, the proportion of identical responses for causation and blame were 78%.

		Between-subjects					Within-subjects				
Contrast	*Variable*	*df*	*t*	*p*	Cohen's *d*	95% CI	*df*	*t*	*p*	Cohen's *d*	95% CI
NN v. PN	Causation Primary	104	–9.52	<.001	1.85	[–2.30;–1.3	123	–8.79	<.001	0.79	[–.99;–.59]
	Knowledge	94	–2.49	0.015	0.48	[–.87;–.10]	123	–4.77	<.001	0.43	[–.61;–.24]
	Desire	104	0.21	0.834	0.04	[–.34;.42]	123	–1.91	0.058	0.17	[–.39;–.01]
	Blame	104	–11.99	<.001	2.33	[–2.81;–1.8	123	–10.33	<.001	0.93	[–1.14;–.76]
NN v. NP	Causation Primary	92	–5.45	<.001	1.08	[–1.50;–.67]	115	–4.33	<.001	0.40	[–.59;–.21]
	Knowledge	101	–.64	0.525	0.13	[–.51;.26]	115	–.63	0.529	0.06	[–.24;.12]
	Desire	101	0.71	0.478	0.14	[–.25;.53]	115	–.58	0.566	0.05	[–.24;.13]
	Blame	86	–6.29	<.001	1.25	[–1.67;–.83]	115	–4.05	<.001	0.38	[–.56;–.19]
NN v. SN	Causation Primary	107	–4.64	<.001	0.89	[–1.28;–.49]	113	–1.34	0.184	0.13	[–.31;.06]
	Knowledge	106	–1.70	0.092	0.33	[–.70;.05]	113	0.00	1.000	0.00	[–.18;.18]
	Desire	107	–.63	0.531	0.12	[–.50;.26]	113	–.20	0.842	0.02	[–.20;.17]
	Blame	107	–5.59	<.001	1.07	[–1.47;–.67]	113	–1.93	0.056	0.18	[–.37;.004]

Table 2: Comparison of effect sizes for the no norm v. pertinent norm (NN v. PN), nonpertinent norm (NN v. NP), and silly norm (NN v. SN) conditions across ascriptions of causation, mental states, and blame. 95% confidence intervals for the reported *d*-values.

4.3.3 Meta-analysis of effects across designs for all three experiments

In the within-subjects designs, the vast majority of participants did *not* perceive a difference in causality due to the violation of nonpertinent or silly norms as compared to the no norm condition. In order to provide statistical support for the claim that the effect sizes in the within-subjects design were significantly smaller than the effect sizes in the between-subjects design, we ran meta-analyses contrasting the results of the two design types from Experiments 1–3. Figure 6 presents the mean effects of norm status on all DVs for all three contrasting pairs, estimated with the restricted maximum-likelihood method based on a random effects model (see Viechtbauer, 2010). As shown, the effect of norm status on causation and blame was significantly and substantially reduced across all three contrastive conditions (pertinent norm, nonpertinent norm, silly norm) in the within-subjects design. For example, the effect sizes on causation and blame in the no norm v. pertinent norm contrast were reduced from $d = 1.44$ and $d = 1.66$ in the between-subjects conditions to $d = 0.78$ and $d = 0.91$ in the within-subjects conditions, respectively. Even more starkly, in the nonpertinent and silly norm conditions, the moderate-to-large sized effects on causation and blame in the between-subjects conditions (nonpertinent norm: $d = 0.82$, $d = 0.88$; silly norm: $d = 0.68$, $d = 0.79$) were all reduced to very small effects in the within-subjects conditions (nonpertinent norm: $d = 0.16$, $d = 0.13$; silly norm: $d = 0.25$, $d = 0.04$).

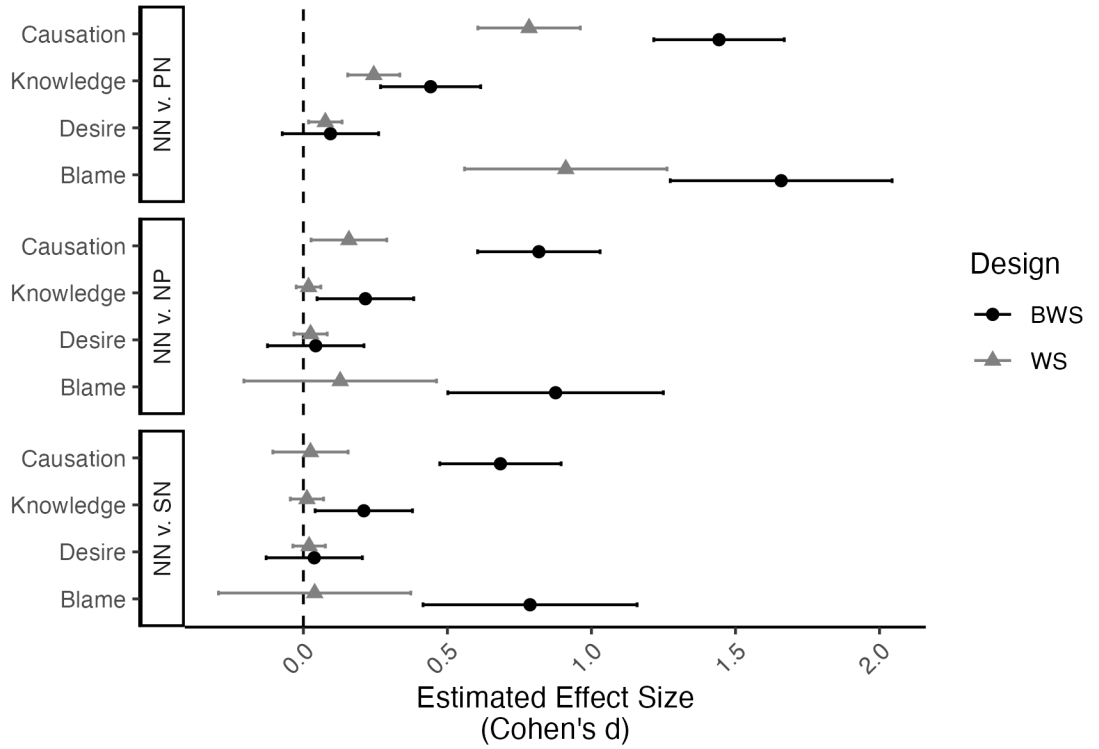

Figure 6: Effects of norm status on the dependent variables across designs in a random effects model in terms of Cohen's *d*. Error bars denote 95% confidence intervals.

4.3.4 Mediation Analyses for Experiments 1 and 3

We ran a series of mediation analyses on the combined between-subjects data from the structurally very similar Experiments 1 and 3, distinguishing between the three contrastive pairs of interest: Pertinent norm, nonpertinent norm and silly norm, each contrasted with the no norm condition. In each of the three analyses, the significant total effect of norm status on causation was reduced to non-significance, once blame was taken into account as a mediator (norm v. pertinent norm: c = 2.35, $p < .001$, c' = .32, $p > .05$; norm v. nonpertinent norm, c = .83, $p < .001$, c' = .10, $p > .05$, norm v. silly norm: c = .42, $p <. 001$, c' = .05, $p >.05$; see Appendix Section 3.3.1, Figures 3a–c). For none of the three norm contrasts, desire and knowledge proved significant mediators (all *a* paths nonsignificant, *p*s < .001 for all remaining direct effects or c' paths; Appendix Section 3.3.1, Figures 4a–c). Just as the Bias View would predict, there is *no* significant direct effect of norm type on causation. The entire effect is mediated – and, on this view, distorted – by blame. The Responsibility View can accommodate this pattern for the no norm v. pertinent norm contrast: The norm has a justifiable influence on blame/responsibility and hence on attributed causation. However, for the nonpertinent and silly norm conditions, the results are problematic, as norm status should *not* influence blame or responsibility, and hence neither should it affect perceived causation.

4.4 Discussion

All our results from Experiments 1 and 2 replicated. Despite their normative absurdity and irrelevance to the outcome at hand, nonpertinent and silly norms had a large effect on perceived causality, blame and the other moral DVs. This, we take it, constitutes a serious problem for the Responsibility View. Pointing to an indirect effect via the mediators knowledge and desire is not an option as their difference *vis-à-vis* the no norm condition was either nonsignificant or very limited in size. Importantly, the effect of the silly norm disappears entirely in the within-subjects design, and, for the nonpertinent norm is drastically reduced from a large effect (d = 1.08) in the between-subjects design to a small one in the within-subjects design (d = .40), and driven by a minority of participants (about 20%, the rest judge the two cases identically). This shows that, in conditions that encourage reflective judgment (having the two cases side-by-side), the vast majority of people do *not* seem to view such norms as relevant to causal judgment (and the same, by and large, holds for blame and responsibility). In other words, when participants are given the opportunity to make a comparative assessment of the norm-adhering and nonpertinent or silly norm-violating agents, most of them appear to judge them identically with respect to their causal contribution, moral responsibility, and blameworthiness.

As regards the pertinent norm: The very large effects measured in the between-subjects design on causation (d = 1.85) and blame (d = 2.33) are significantly reduced to about half in the within-subjects designs (d = .79, d = .93), but remain large. Given the significant reduction, their extraordinary between-subjects size is presumably at least partially driven by bias. However, the within-subjects effects and the fact that about 60% of participants rate causation differently in the no norm v. pertinent norm contrast suggest that the majority of people *do* think that pertinent norms are relevant to the assessment of causation (and the same holds for blame). It might thus be the case that the (pertinent) Norm Effect, does not constitute a bias – and the Bias View, with its emphatic commitment to the descriptive nature of causality faces a challenge here.

Perhaps (though we do not want to commit to any particular solution here), advocates of the Bias View would respond to the challenge by pointing out that certain biases are extraordinarily *sticky*. Kahnemann and Tversky (1984), in one of their landmark papers on framing effects, for instance, found the formulation of risky options ("lives saved" v. "lives lost") to have a strong impact on choice. Although a bias as clear as they come, people continued to find appeal in distinct answers after having been confronted with their inconsistent choices. "In their stubborn appeal," Kahneman and Tversky write, "framing effects resemble perceptual illusions more than computational errors." (1984, p. 5). Perhaps the stickiness of the (pertinent) Norm Effect in the within-subjects design is – for those who are subject to it – not unlike the continued and puzzling effect of framing on choices.

The Responsibility View of ordinary causation assigns a special role to morally relevant mediators. Consequently, advocates of this view might argue that that the mediators we have tested thus far are not the most appropriate ones. Since our vignettes involve accidents, they might argue, it comes as no surprise that participants do not ascribe knowledge or desire to the agent. And indeed, mean knowledge and desire attribution are extremely low in both our experiments with the Festival vignette (Experiments 1 and 2, all Ms < 2.50, significantly below the midpoint of the scale, all *ps* < .001), and those with the Trash Bag vignette (Experiment 3, all Ms < 2.24, significantly below the midpoint, all *ps* < .001). When it comes to accidents, the more appropriate mediator may be the carelessness or negligence of the agent, which is standardly determined in relation to how reasonably foreseeable the accident was (Engelmann & Waldmann, 2021, 2022; Kirfel & Lagnado, 2021a, 2021b; Lagnado & Channon, 2008; Jaeger, 2023; Kneer & Machery, 2019; Kneer, 2022; Nobes & Martin, 2022; Sarin & Cushman, 2023, 2024, Murray et al. 2023, Tobia, 2018). It could turn out that participants judge agents that violate norms – even nonpertinent or silly ones – as acting more negligently than their norm-adhering counterparts and thus rightfully consider them more responsible. In Experiments 4 and 5 we explore whether the Responsibility View can be saved by recourse to foreseeability as an alternative mediator.

5. Experiment 4

Experiment 4 investigates whether the findings of the previous experiments can be explained by recourse to a different potential mediator, namely the foreseeability of an accident. Quite independently of the narrower concerns of the academic debate on the Norm Effect, this question is of central legal relevance, as the foreseeability of an outcome is crucial in assessing causation in the law: on the legal view, an agent can only be held liable for a harmful outcome if said agent could have foreseen its coming about (Dressler, 2015; Goldberg & Zipursky, 2010; Owen, 2009). Furthermore, whether or not an accident was reasonably foreseeable is a key desideratum in the attribution of negligence and recklessness, two legally inculpating mental states.

It is noteworthy that from a legal perspective, nonpertinent and silly norm violations should not impact foreseeability (for discussion, see *e.g.* Brown, 2023; Margoni & Brown, 2023; Green, 1961; VerSteeg, 2011). The reasons for this are twofold. First, the law only cares about norm violations that stand in the appropriate relation to the outcome, and nonpertinent and silly norm-violations evidently do not. Second, the law is in the business of making an objective *ex ante* assessment of whether a certain outcome was foreseeable or not, and is explicit to exclude factors irrelevant to the outcome from its consideration. Thus, from a legal perspective, not only would an unmediated effect of nonpertinent and silly norm-violations on causation and responsibility constitute a clear bias, but an effect of nonpertinent and silly norms on *foreseeability* would be problematic as well. Differently put, if, as the Responsibility View predicts, the effect of nonpertinent and silly norms can be explained by recourse to foreseeability, and foreseeability judgements are sensitive to the kind of norm being violated, then that finding would be legally problematic in its own right.

In what follows, we sought to explore this hypothesis. In order to account for the hindsight bias which frequently besets foreseeability judgements (Kamin & Rachlinski, 1995; Margoni & Surian, 2022; Kneer & Skoczen, 2023; Margoni & Brown, 2023; Rachlinski, 1998, 2000; for a review, see Roese & Vohs, 2012), we presented participants with both *ex ante* (outcome information not yet available) and *ex post* (outcome information available) conditions of the Trash Bag vignette. Although the law is interested in an *ex ante* assessment of foreseeability, it most frequently operates in an *ex post* context. By running Experiments 4 and 5 both in *ex ante* and *ex post* presentation order, we hoped to gain greater insight into whether foreseeability could be a potential mediator, as proponents of the Responsibility View might suggest, and whether judgements of foreseeability themselves are biased (*e.g.* due to a hindsight bias).

5.1 Participants

We recruited 1014 participants on Prolific. Their IP address was restricted to the United States. In line with our preregistration criteria,[12] we excluded participants who failed a general attention check, spent less than 15 seconds reading the vignette, or were not native English speakers. 960 participants remained (exclusion rate: 5.3%; female: 46%; mean age: 42 years, SD = 13 years, range = 19–94 years).

5.2 Methods and materials

The study took a 4 (norm type: no norm v. pertinent norm v. nonpertinent norm v. silly norm) × 2 (presentation of foreseeability question: *ex ante* v. *ex post*) between-subjects design. Participants were randomly assigned to one of the four conditions of the Trash bag vignette from Experiment 3.

Participants in the *ex post* conditions were given the vignette in full (*i.e.* including the building catching fire), and asked the exact same questions as in Experiment 3, except that we replaced the foreknowledge and desire questions with a single foreseeability question. They were asked to rate the extent to which they agree or disagree with the following statement (label in bold omitted):

> **Foreseeability:** "Mark could have reasonably foreseen the coming about of injuries." (1 = completely disagree, 7 = completely agree)

Participants in the *ex ante* conditions were given the vignette only up to the mention of Mark placing his trash bags outside and were asked to make an initial evaluation as to the foreseeability

[12] Available under https://aspredicted.org/BPB_3YS.

of an accident. Afterwards, participants were told about the accident and asked to rate the causal contributions of the primary and secondary agent (Mark and the workers) as well as assess the moral variables of blame, responsibility, and punishment. All responses were collected on 7-point Likert scales.

5.3 Results

ANOVAs. We ran a series of 4 (norm type) × 2 (presentation order) between-subjects ANOVAs for all dependent variables. As regards foreseeability, we found a significant and moderately-sized main effect of norm type ($p < .001$, $\eta_p^2 = .079$) as well as a large effect of presentation order ($p < .001$, $\eta_p^2 = .162$). The interaction was nonsignificant ($p = .170$, $\eta_p^2 = .005$). For causation, we found a significant and large effect of norm type ($p < .001$, $\eta_p^2 = .165$), though neither presentation order ($p = .105$, $\eta_p^2 = .003$), nor the interaction ($p = .400$, $\eta_p^2 = .003$) seemed to have influenced participants' judgements. For our moral variables of blame, responsibility, and punishment, we found a large effect of norm type throughout (blame: $p < .001$, $\eta_p^2 = .177$; responsibility: $p < .001$, $\eta_p^2 = .166$; punishment: $p < .001$, $\eta_p^2 = .199$), while both the presentation order and interaction remained nonsignificant (all *ps* > .174) (see Appendix Section 3.5 for full results).

		Ex ante					Ex post				
Contrast	*Variable*	*df*	*t*	*p*	Cohen's *d*	95% CI	*df*	*t*	*p*	Cohen's *d*	95% CI
NN v. PN	Causation Primary	242	–9.58	<.001	1.25	[–1.52;–.97]	182	–11.53	<.001	1.58	[–1.89;–1.27]
	Foreseeability	242	–6.05	<.001	0.79	[–1.05;–.52]	210	–7.27	<.001	1.00	[–1.28;–.71]
	Blame	229	–10.57	<.001	1.35	[–1.63;–1.07]	196	–11.52	<.001	1.58	[–1.89;–1.27]
NN v. NP	Causation Primary	285	–6.04	<.001	0.71	[–.95;–.47]	216	–4.33	<.001	0.59	[–.86;–.32]
	Foreseeability	285	–.24	0.815	0.03	[–.26;.20]	216	–2.54	0.012	0.34	[–.61;–.08]
	Blame	285	–6.74	<.001	0.80	[–1.04;–.56]	216	–4.42	<.001	0.60	[–.87;–.33]
NN v. SN	Causation Primary	281	–6.44	<.001	0.77	[–1.01;–.52]	216	–4.34	<.001	0.59	[–.86;–.32]
	Foreseeability	281	–.51	0.612	0.06	[–.29;.17]	216	–2.03	0.044	0.28	[–.54;–.01]
	Blame	281	–6.30	<.001	0.75	[–.99;–.51]	216	–4.66	<.001	0.63	[–.90;–.36]

Table 3: Comparison of effect sizes for the no norm v. pertinent norm (NN v. PN), nonpertinent norm (NN v. NP), and silly norm (NN v. SN) conditions across ascriptions of causation, mental states, and blame. 95% confidence intervals for the reported *d*-values.

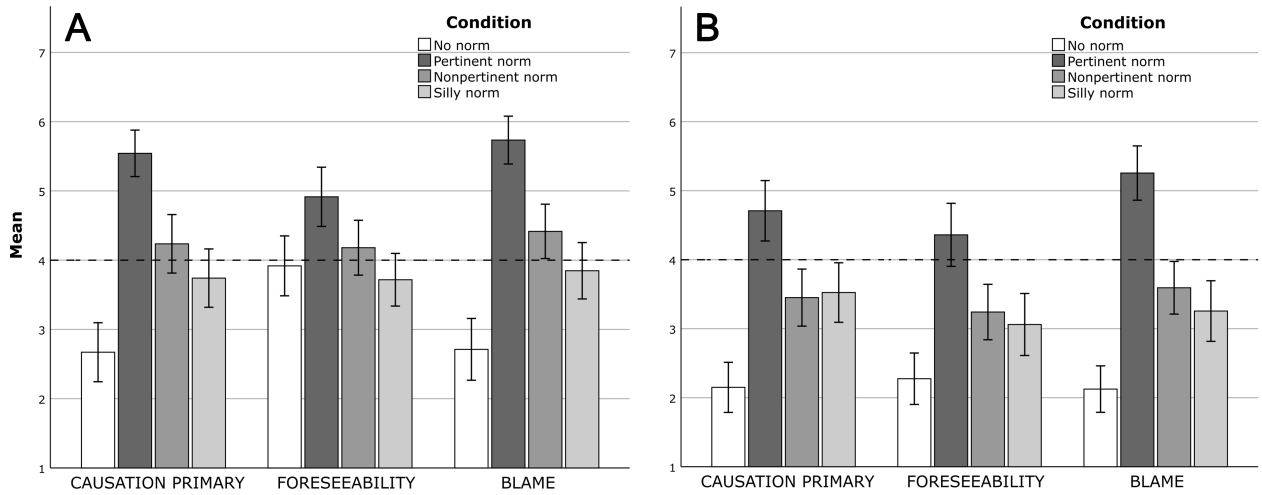

Figure 7: Comparison of means for key dependent variables in the *ex ante* (A) and *ex post* (B) conditions. Error bars denote 95% confidence intervals.

No norm v. pertinent norm. In comparing the no norm and pertinent norm conditions, we found significant differences in foreseeability, Mark's causal contribution, as well as the moral variables across both *ex ante* and *ex post* presentation order (all *p*s < .001). The effect sizes in the *ex post* condition were considerably larger (all *d*s > 1.00) than in the *ex ante* condition (all *d*s > .78). This reduction in effect size is consistent with recent work on the hindsight bias afflicting *mens rea* attribution and related variables (see Kneer & Skoczen, 2023).

No norm v. nonpertinent norm. In the *ex post* condition, participants judged Mark's causal contribution as well as the moral variables to differ significantly across norm type (all *p*s < .001, all *d*s > .56). Given that the effect of norm type on foreseeability was significant ($p = .012$, $d = .34$), one might consider it as a mediator that renders the effect of nonpertinent norm-infractions on causation plausible according to the logic of the Responsibility View. However, as our *ex ante* results show, this might be too quick. In the *ex ante* conditions, we again found a significant effect of norm type on causation and the moral DVs (all *p*s < .001, all *d*s > .69). Crucially, however, we did *not* find a difference in foreseeability ($p = .815$) suggesting that, as long as one avoids distortion due to the hindsight bias, recourse to foreseeability *cannot* rehabilitate the Responsibility View.

No norm v. silly norm. The results for the no norm v. silly norm contrast replicate the pattern reported in the previous paragraph. Whereas participants in the *ex post* conditions judged different all aforementioned dependent variables (all *p*s < .001, all *d*s > .58), including, though just about, the foreseeability of an accident ($p = .044$, $d = .28$), participants in the *ex ante* conditions perceived a difference only with respect to Mark's causal contribution and the moral variables (all *p*s < .001, all *d*s > .69) but *not* the foreseeability of the accident ($p = .612$).

5.4 Discussion

Consistent with the results of Experiments 1–3, the effects of the *pertinent* norm on causation and the moral variables were significant and very pronounced (all *ds* > 1.18). In the pertinent norm condition, the effect on foreseeability, too – even when assessed *ex ante* – was significant and close-to-large ($d = .79$). This finding is consistent with predictions made by the Responsibility View, according to which the influence of *prima facie* irrelevant factors such as norm violation on causal responsibility can be explained by aid of a mediator such as foreseeability: pertinent norm violations, one might argue, *should* impact foreseeability, and thereby moral responsibility and blame. Given the tight connection between moral and causal responsibility, and the fact that the norm violation stands in relation to the outcome, their impact on perceived causal responsibility is explained. Indeed, these findings are consistent with the Counterfactual View as well: since the norm violation in question was pertinent to the outcome at hand, the results can also be explained by recourse to the participants' being drawn to salient counterfactuals in which, absent the norm violation, the outcome failed to obtain.

For the Responsibility View, things are considerably more problematic as regards nonpertinent and silly norms. Replicating the findings from Experiments 1–3, we again found a significant and considerable impact on perceived causation and the moral variables (all *ds* > .58). Furthermore, when the reasonable foreseeability of an accident was assessed *ex ante*, it did not prove sensitive to nonpertinent and silly norm violations. For nonpertinent and silly norms, then, attempts to rehabilitate the Responsibility View by aid of plausible mediators such as foreseeability, knowledge and desire have thus far all failed.

Going beyond the debate on causation, the fact that *ex post* assessments of foreseeability are responsive to norm violations is an interested, novel and worrisome finding: in the law, juries are to judge foreseeability with respect to the agent's circumstances and epistemic situation (*i.e.* in an *ex ante* fashion). As our results show, the hindsight bias might make this difficult, just as it distorts a whole range of other variables relevant to negligence attribution (see Kneer & Machery, 2019; Kneer & Skoczen, 2023).

6. Experiment 5

Our results so far are troubling for the Responsibility View: we have found that nonpertinent and silly norm impact causation, and that the third possible mediator – foreseeability – is of no help to explain this effect. To explore whether these findings generalise beyond the Trash Bag scenario, we replicated the experiment with the novel Shooting Range vignette (full scenario in Appendix Section 4.3).

6.1 Participants

1034 participants were recruited online on Amazon Mechanical Turk. Their IP address was restricted to the United States. As preregistered,[13] we excluded participants who failed a general attention check, spent less than 10 seconds on the page presenting the vignette, or were not native English speakers. 680 participants remained (exclusion rate: 34.2%; female: 49%; mean age: 42 years, SD = 13 years, range = 20–94 years).

6.2 Methods and materials

Just like Experiment 4, the study took a 4 (norm type: no norm v. pertinent norm v. nonpertinent norm v. silly norm) × 2 (presentation of foreseeability question: *ex ante* v. *ex post*) between-subjects design. Participants were randomly assigned to one of the eight conditions of the Shooting range vignette. The story has Mark shooting at an outdoor shooting range while Lauren is hiking in the nearby forest. The sudden appearance of a wild boar frightens Lauren, who tumbles down a hill and comes to halt right in front of the bullet Mark shot moments earlier. The bullet lodges itself in her leg and Lauren has to be taken to the hospital.

The no norm condition mentions a shooting range in regular operation. In the pertinent norm condition, Mark practices at the shooting range although it's closed. In the nonpertinent norm condition, it is prohibited to use the shooting range unless one wears protective gloves and glasses, and Mark does not wear any. In the silly norm condition, it is forbidden to bring any type of food or drinks to the shooting range, and Mark sneaks in a bag of potato chips and a soft drink.

As in Experiment 4, participants in the *ex post* conditions were given the vignette in full (*i.e.* including the injury of Lauren), while participants in the *ex ante* conditions were given the vignette only up to the mention of Lauren hiking. All participants had to rate the causal contributions of Mark and the boar, the foreseeability of the accident, as well as the moral variables of blame, responsibility, and punishment. In the *ex post* conditions, the foreseeability question came after the causal questions, whereas in the *ex ante* conditions, it came before the outcome was revealed. The questions were phrased as in the experiments above, and responses were recorded on 7-point Likert scales.

6.3 Results

ANOVAs. A 4 (norm type) × 2 (presentation order) between-subjects ANOVA revealed a significant main effect of both order and norm type on foreseeability (both *p*s < .001), with a small-

[13] Available under https://aspredicted.org/B7H_QXS.

to-moderate effect size for order ($\eta_p^2 = .057$) and a moderate effect size for norm type ($\eta_p^2 = .084$). The interaction was close to significant ($p = .053$). The main effect of norm type on Mark's perceived causal contribution was significant and large ($p < .001$, $\eta_p^2 = .204$) and was accompanied by a significant yet small effect of presentation order ($p < .001$, $\eta_p^2 = .024$). We further found significant and large main effects of norm type on all moral variables (all ps $< .001$, all η_p^2s $> .256$), and small main effects for presentation order (all ps $< .007$, all η_p^2s $< .029$).

		Ex ante					Ex post				
Contrast	*Variable*	*df*	*t*	*p*	Cohen's *d*	95% CI	*df*	*t*	*p*	Cohen's *d*	95% CI
NN v. PN	Causation Primary	165	–10.69	<.001	1.67	[–2.02;–1.31]	160	–8.86	<.001	1.38	[–1.72;–1.04]
	Foreseeability	165	–3.21	0.002	0.50	[–.81;–.19]	160	–7.03	<.001	1.08	[–1.41;–.76]
	Blame	145	–10.67	<.001	1.69	[–2.04;–1.33]	164	–11.94	<.001	1.86	[–2.22;–1.49]
NN v. NP	Causation Primary	160	–5.14	<.001	0.81	[–1.13;–.49]	168	–4.70	<.001	0.71	[–1.02;–.40]
	Foreseeability	160	–.88	0.376	0.14	[–.45;.17]	169	–3.48	0.001	0.53	[–.84;–.23]
	Blame	160	–5.72	<.001	0.90	[–1.23;–.58]	168	–5.74	<.001	0.87	[–1.18;–.55]
NN v. SN	Causation Primary	156	–3.54	0.001	0.57	[–.88;–.25]	156	–4.85	<.001	0.76	[–1.08;–.44]
	Foreseeability	156	0.70	0.488	0.11	[–.20;.42]	155	–2.68	0.008	0.42	[–.73;–.11]
	Blame	156	–3.75	<.001	0.60	[–.92;–.28]	151	–4.07	<.001	0.64	[–.95;–.32]

Table 4: Comparison of effect sizes for the no norm v. pertinent norm (NN v. PN), nonpertinent norm (NN v. NP), and silly norm (NN v. SN) conditions across ascriptions of causation, mental states, and blame. 95% confidence intervals for the reported *d*-values.

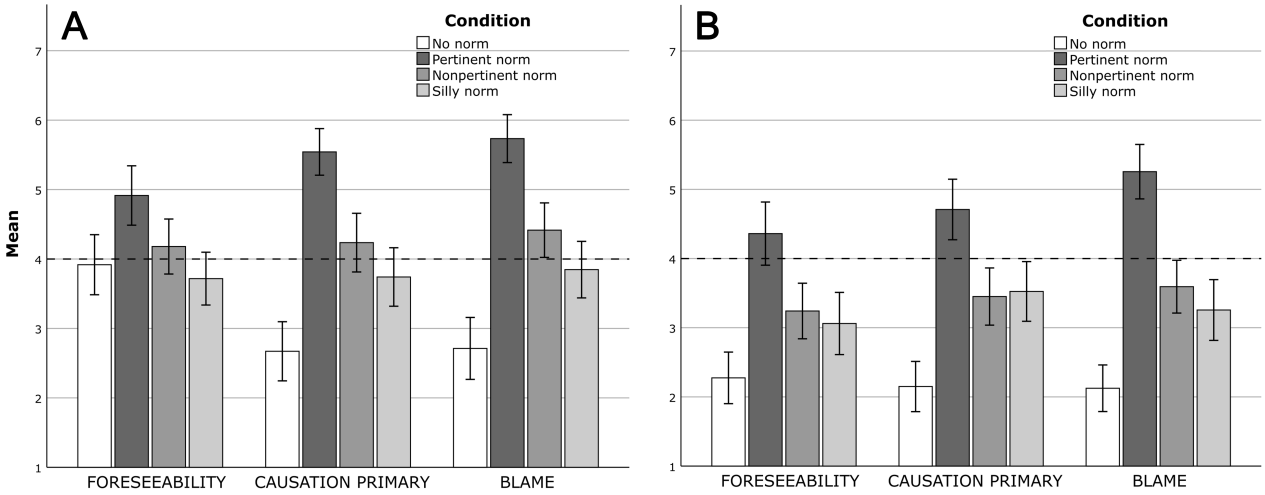

Figure 8: Comparison of means of the core dependent variables in the *ex ante* (A) and *ex post* (B) conditions. Error bars denote 95% confidence intervals.

No norm v. pertinent norm. A comparison between the no norm and pertinent norm conditions yielded significant differences for all variables in both *ex ante* and *ex post* presentation order (all ps $< .003$). For causation and the moral DVs, the effect sizes were large in both designs (all $ds > 1.37$). For foreseeability, the effect size was large in the *ex post* condition ($d = 1.08$), though

significantly smaller in the *ex ante* condition (d = .50). Given this discrepancy, it is doubtful that the very large effect of norm status on causation (d = 1.67) in the *ex ante* design can exhaustively be explained by the only mid-sized effect on foreseeability.

No norm v. nonpertinent norm. In the *ex post* conditions, there was a significant effect on causation, the moral variables and foreseeability (all ps < .001, all ds > .52). In the *ex ante* conditions, we also found significant (all ps < .001) and sizeable effects on causation (d = .81) and the moral DVs (all ds > .84). Importantly, however, there was no significant effect of norm status on foreseeability (p = .376), suggesting that the significant and large norm effect on causation is not mediated by foreseeability once the hindsight bias is corrected for.

No norm v. silly norm. In the *ex post* condition, we again found a significant norm effect on causation, foreseeability and the moral variables (all ps < .001, all ds > .41). In the *ex ante* condition, however, foreseeability again proved nonsignificant (p = .488), which suggests that it cannot explain the significant norm effect on causation (p < .001, d = .57) or the moral variables (all ps < .01, all ds > .41).

6.4 Discussion

Experiment 5, beyond replicating all key findings of Experiment 4, puts a little more pressure on the Responsibility View in the no norm v. pertinent norm contrast. The large norm effect on causation (d = 1.67) in the *ex ante* condition cannot be exhaustively explained by the significant, though only mid-sized impact of norm status on foreseeability (d = .50). Hence, a considerable residual effect of norm status on perceived causation is unaccounted for. And, once again, the influence of nonpertinent and silly norm-violations on perceived causation in the *ex ante* conditions cannot be explained by recourse to foreseeability, because here, too, the latter was nonsignificant.

Finally, the drastic difference in foreseeability judgements *ex post* vis-à-vis *ex ante* (up to a Cohen's d of .50 difference) points to the hindsight bias: the tendency for an event to be deemed more predicable or probable after one has learned that the event has in fact occurred. Thus, although it seemed like foreseeability would act as a suitable mediator for responsibility – and by extension causation – judgements in the *ex post* conditions, the foreseeability judgements themselves are due to hindsight bias and thus can do little to render the Responsibility View more plausible.

7. General Discussion

In this section, we will briefly discuss the implications of all our results in somewhat more detail. Given the considerable ground the studies have covered, Table 5 provides a brief overview of our aims and findings.

Experiment	Vignette	Design	Aims	Findings
#1	Festival	BWS	Show that pertinent, nonpertinent, and silly norm violations impact causation without being mediated by foreknowledge and desire	Medium-to-large effects of the three norm violation conditions on causation, while foreknowledge and desire proved nonsignificant
#2	Festival	WS	Show that the impact of pertinent, nonpertinent, and silly norm violations on causation is, upon reflection by participants, reduced or eliminated entirely	Effect size of pertinent norm violations on causation significantly reduced from large to medium; effects of nonpertinent and silly norms on causation eliminated entirely
#3	Trash Bag	BWS & WS	Replicate the findings of Experiments 1 & 2 with a novel vignette	Succesful replication of all results; the effect size reduction in this vignette is even more drastic (see the meta-analysis in Figure 6)
#4	Trash Bag	BWS (*ex ante* v. *ex post*)	(1) Show that pertinent, nonpertinent, and silly norm violations impact causation without being mediated by foreseeability (negligence); (2) explore hindsight bias on ex post causal judgements	(1) Foreseeability can only partially explain the effect of pertinent norm violation on causation, and cannot explain the effect of nonpertinent or silly norm violations; (2) Foreseeability and causal judgements manifest hindsight bias
#5	Shooting Range	BWS (*ex ante* v. *ex post*)	Replicate the findings of Experiment 4 with a novel vignette	Succesful replication of findings (1) and (2).

Table 5: Overview of the five experiments, their aims, and their key findings.

7.1 Implications for the Responsibility and Bias Views

According to the Responsibility View, the ordinary concept of causation is strongly intertwined with moral responsibility. On this account, factors which legitimately increase the attribution of moral responsibility, such as the foreknowledge of harm or the agent's desire to harm, can be viewed as legitimately increasing perceived causation. The Bias View, by contrast, takes the concept of causation to be nonnormative. According to its proponents, cases where moral factors increase perceived causation testify to a performance error of human judgment: people are inclined to blame an agent who causes harm more than one who doesn't and, in an attempt of post-hoc rationalization, exaggerate her causal contribution.

Advocates of the Responsibility View acknowledge that not just any factor that *could* influence perceived moral responsibility *should* influence perceived causality. Only factors that are legitimately connected to moral responsibility proper are viewed as exerting a warranted impact on perceived causality. This excludes, for instance, the agent's race, gender, or moral character. It also excludes the violations of nonpertinent or silly norms, *i.e.* norms which are unrelated to a

specific action's outcome. Importantly, there are exceptions: if certain features that are *prima facie* irrelevant to moral responsibility, such as general moral character, engender reasonable inferences to factors which *are* relevant (such as *e.g.* the mental state of the agent), advocates of the Responsibility View argue, this should not be considered as evidence against the account.

We have explored two challenges to the Responsibility View. First, we have shown that the violation of nonpertinent and silly norms unconnected to the resulting harm have a significant and considerable impact on perceived causality, with medium to large effect sizes. According to all aforementioned views, they should *not* impact moral responsibility or blame. However, they do, and – in line with the Bias View – presumably thereby influence perceived causation. Given that potential, reasonable mediators (foreknowledge, desire to harm, foreseeability) of interest did not prove significant, it is difficult for the Responsibility View to tell a convincing story here. What is more, in a within-subjects design (Experiments 2 and 3) we show that participants *themselves* hold that nonpertinent and silly norms should *not* influence causality attributions: the vast majority of them rated the causal impact (and blame) of the norm-abiding and norm-violating identically. Grist to the mill of the Bias View.

Replicating extant findings, we also found a powerful effect of norm-violations pertinent to the action. Proponents of the Responsibility View seem to hold that pertinent norms should *only* exert an influence on perceived causation if it were mediated by reasonable inferences regarding legitimate influences on moral responsibility. However, and this constitutes our second challenge, the large effects (Cohen's $ds > 1.00$) we found for causation in between-subjects designs cannot *exhaustively* be explained through inferences regarding *mens rea* (foresight, foreseeability, desire). As Experiments 4 and 5 demonstrate, however, at least foreseeability seems able to *partially* account for the effect.

We agree with Sytsma's warning that "researchers need to carefully consider and control for the inferences that participants might draw concerning the agents' mental states and motivations" (2019b, p. 25). Our Experiments 4 and 5 address this point further. Scholars who suggest that moral responsibility and causation are driven by a particular inference to *mens rea*, such as negligence (*i.e.* reasonable foreseeability), must be careful to distinguish when such an inference is warranted, and when it is not. As the results of Experiment 5 show, the large effect of pertinent norm violations on foreseeability *ex post* ($d = 1.08$) which seems to explain the large effect on causation ($d = 1.38$), shrinks to less than half ($d = .50$) once the hindsight bias is controlled for, and can no longer fully explain the very large effect on perceived causation ($d = 1.67$).

Although the data reported favours the Bias View, in particular as regards the effects exerted by nonpertinent and silly norms, this does not yet mean that the (pertinent) Norm Effect on causality attributions itself constitutes a bias. After all, one might formulate a weaker version of the Responsibility View, according to which the violation of pertinent norms, *even if unmediated by*

other factors such as mens rea, exerts a *legitimate* influence on moral responsibility and (therefore) attributed causality. An account of this sort need not collapse into the unattractive Anything-Goes View, as long as moral-philosophical reasons are provided why norm-infractions are relevant to the responsibility of the agent – and such reasons do not seem that hard to come by. One interesting data point in favour of this more permissive version of the Responsibility View is provided by the within-subjects results: in contrast to the nonpertinent and silly norm comparisons, about two-thirds of our participants *did* consider the violation of pertinent norms relevant to the assessment of responsibility and causation. Note that this part of the finding could constitute a challenge to the Bias View, which lobbies for an unequivocally descriptive concept of causation. One potential reply, we have argued, might consist in the *stickiness* of certain kinds of biases: On any reasonable account, the framing effects arising from a negative or positive formulation of casualties ("lives lost" v. "lives saved") do constitute a bias, and yet, as Kahnemann and Tversky (1984) reported, a considerable proportion of people failed to view them as such and remedy their responses when given a chance.

7.2 Implications for the Pragmatic View

As regards the cases at hand, the predictions of the Pragmatic View are largely coextensive with those of the Bias View and the Anything-Goes interpretation of the Responsibility View. As long as participants *interpret* the causality question in terms of accountability or moral responsibility, they will respond differently than when interpreting it in purely descriptive terms. Hence the differences in causality attributions across the no norm v. norm conditions where they arise, and hence the absence of any such difference due to a homogenous, descriptive interpretation in most of the within-subjects conditions.

Although, overall, the Pragmatic View, like the Bias View, fares better than the other accounts, it also faces some difficulties explaining the results in the within-subjects no norm v. pertinent norm contrast: Two thirds of the participants assess the causal contribution of the agents differently. But why, one might justifiably ask, should participants interpret the *same* question in two very different ways when assessing the no norm and the pertinent norm conditions *side-by-side*? Note that participants first read both scenarios, and then responded to a single causality question asking them to assess both agents – one Likert scale directly below the other. In contrast to ordinary within-subjects designs, the stimuli aimed specifically at making the similarities of the two scenarios – and the identity of the question – maximally clear to the reader. So it is not evident what advocates to the Pragmatic View could respond in order to explain this part of the findings. And even though the theoretical precommitments seem smaller, and perhaps the general permissiveness of the Pragmatic View larger than that of the Bias View, the latter, we suggested, at least has *some* explanation at offer for these results, namely the potential *stickiness* of bias with some subjects, which have long been documented to arise in other areas, such as framing effects, too.

7.3 Implications for the Counterfactual View

Recall that the general idea underlying the Counterfactual View is that participants who are confronted with a norm violation are drawn to consider the counterfactual in which said norm was *not* violated. Noticing that in such a case the outcome would not obtain, they think of the agent as the cause of the outcome. We have fleshed this out more precisely by considering a particular version of the Counterfactual View, namely the Necessity-Sufficiency Model proposed by Icard et al. (2017). As described above, the Necessity-Sufficiency model holds that the Norm Effect is the result of norm violations affecting our perception of the necessity or sufficiency of certain causal factors. This is because norm violations are standardly regarded as abnormal causal factors and, on this model, judged on the basis of their necessity for the outcome at hand. How does this prediction square with our findings?

As it turns out, the model proposed by Icard and colleagues is consistent with our findings as regards *pertinent* norm violations. In Experiment 1, for instance, we find a strong influence of pertinent norm violations on causal ascriptions (Cohen's d = 1.01), and although Experiment 2 reveals this effect to be partially driven by bias, a considerable residual effect remains (d = 0.70). The Necessity-Sufficiency Model is in a good position to explain this residual effect. After all, Mark's norm-deviant behaviour (smoking despite the prohibition) is directly linked to the harmful outcome at hand (the explosion). If participants were to consider the counterfactual in which Mark did not smoke, they would come to see that it was *necessary* for the outcome to obtain. Thus, the Necessity-Sufficiency Model would predict heightened ascriptions of causation for Mark – which is precisely what we find.

Nevertheless, the Counterfactual View, too, struggles to accommodate the between-subjects findings of the nonpertinent and silly norm conditions. This is because the nonpertinent and silly norm-violations, unlike the pertinent norm violations, do *not* stand in the correct relation to the outcome at hand. Thus, if participants were to consider the counterfactual in which Mark had *not* violated a nonpertinent norm (for instance, if he *did* wear a t-shirt in the Festival scenario), then the harmful outcome would still have obtained. On the Necessity-Sufficiency Model, participants should notice that the norm violation is not necessary for the outcome, and thus not heighten their causal ascriptions towards Mark. This is, however, not what we find. Quite the opposite: despite the patent irrelevance or silliness of the norm violation at hand, participants did judge the agent as more causal. Throughout the paper, we have argued that the Bias View is best positioned to explain this: our findings suggest that participants blame the norm-violating agent, irrespective of the kind of norm violated, and that they, in backwards-rationalising these blame judgements, heighten their causal ascriptions of the norm-violating agent.

It is important to note, however, that our nonpertinent and silly norm findings provide only *preliminary* evidence against the Counterfactual View. This is because the Counterfactual View

is, ultimately, concerned with the *perceived* counterfactual relations between the norm violation and the outcome, and not the counterfactual relations as they actually obtain. While it is clear that nonpertinent and silly norm-violations do not *in fact* make an outcome more or less likely to come about, we did not explicitly test whether participants *perceive* them to make a difference. One way to test for this would be to rerun our studies and ask participants to rate their agreement with a question of the following sort: "Do you think that if Mark had not [insert norm-violating behaviour here], the outcome would have been less likely to come about?"[14] Naturally, if participants were to voice their agreement with this question, that would put their causal competencies into question, and raise a whole host of different issues. Nevertheless, it would be interesting to see whether a question of this sort could mediate some of the effect we find in the nonpertinent and silly norm conditions.

7.4 Implications for the law

Whether or not the pertinent Norm Effect is considered a bias or not, our results demonstrate that attributions of causality are easily influenced by factors that clearly should not play any role. An agent who fails to adhere to some silly norm that happens to be in place should not be judged more causally responsible than one who does. This is not only the view of any reasonable philosopher or moral psychologist, but consistent with the folk view, as the within-subjects data shows. One area where these findings are of great importance is the law: both in torts and criminal law, the *actus reus* (the "guilty act") is one of the two key determinants of liability besides *mens rea* (the "guilty mind"), and in common law jurisdictions (such as the UK and the US), the *actus reus* – or, simply put, causation – is determined by lay juries (Knobe & Shapiro, 2021; Lagnado & Gerstenberg, 2017; Lagnado, 2021; Engelmann & Kirfel, 2024). Furthermore, as Güver & Kneer (2023a) have elaborated, legal practitioners tend to hold that the legal notion of causation *corresponds* to the ordinary notion. Lord Wright, for instance, has stated in a landmark English case that "[c]ausation is to be understood as the man in the street, and not as the scientist or the metaphysician, would understand it."[15] Similarly, Lord Salmon proclaimed that "[w]hat or who caused an event to occur is essentially a practical question of fact which can best be answered by ordinary common sense rather than abstract metaphysical theory". [16] So too the US Supreme Court, which, in the much-cited *Burrage v. United States*, argued that courts should rely on "the common

[14] We would like to suggest, however, that the within-subjects results suggest that the response to this question would be a resounding "no". When presented with the silly/nonpertinent norm condition *and* the no norm condition, participants by and large attribute the *same* degree of causal responsibility to the two agents. The only reasonable explanation of the data is that – upon reflection – participants do *not* consider these norms relevant to causality judgments.

[15] *Yorkshire Dale Steamship Co Ltd v Minister of War Transport* [1942] AC 691 (HL) 706.

[16] *Alphacell Ltd v Woodward* [1972] A.C. 824, 847.

understanding of causation" and explicate causal relations with reference to what it "is natural to say."[17]

If ordinary attributions of causality are easily influenced by bias – as the nonpertinent and silly norm data across between-subjects and within-subjects designs demonstrate – this is problematic for the law: ordinary people, or "blame amateurs", as Alicke calls them, might simply not be capable of keeping morally irrelevant factors at bay, and exaggerate the causal contribution of those whom they are unwarrantedly inclined to blame for harmful outcomes.[18] Norm violations that do not stand in the appropriate relation to the outcome at hand, *i.e.* nonpertinent and silly norm violations, are legally irrelevant, both in common law (see *e.g. Berry v. Sugar Notch Borough,* 43 A. 240 (1899)) and civil law jurisdictions (see *e.g.* Stratenwerth & Kuhlen, 2011). Our findings, on the other hand, suggest that they do factor in, which could result in serious overcriminalization of defendants whose behavior was in some morally or legally irrelevant sense objectionable. This problem is not necessarily limited to common law countries, but might extend to civil law countries, where legal decisions are taken by professional judges: Recent research has shown that legal experts fall prey to the same biases as laypeople, for instance when it comes to outcome bias in *mens rea* attribution (Kneer & Bourgeois-Gironde, 2017; Bourgeois-Gironde & Kneer, 2018, Kneer et al. 2025), confirmation bias (Lidén et al. 2019), sympathy bias (Spamann & Klöhn, 2016; Liu & Li, 2019) or hindsight bias (Strohmaier et al. 2021).

7.5 Limitations and future research

Our studies are limited to three scenarios, three potential mediators, and all our participants are US Americans. For improved external validity, future work should explore a broader range of vignettes as well as other mediators of interest. Moreover, similar experiments should be run across different cultures and languages, in particular non-WEIRD countries (*cf.* Henrich et al. 2010; Henrich, 2020), so as to explore whether the findings constitute a general human disposition of judging causality or not. Some of the cross-cultural work in experimental jurisprudence (see *e.g.* Hannikainen et al. 2021, Hannikainen et al. 2022, Kneer et al. 2025, Tobia et al. 2025) and experimental philosophy (see *e.g.* Knobe, 2023 for a review) has revealed surprising convergence. However, others have documented extensive differences (for a review, see Stich & Machery, 2023).

---

[17] *Burrage v. United States*, 571 US 204 (2014). For further experimental papers concerning causation from a legal perspective, see *e.g.* Solan & Darley (2001, pp. 271–272); Macleod (2019, pp. 982–985); Tobia (2021, pp. 91–92); Summers (2018, pp. 3–5), Prochownik (2022).

[18] Our conclusion here is less radical than the one put forth by Rose (2017), who 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" (p. 1352). For Rose, the folk notion is too unstable and confused to contribute to *any* reasonable account of causation. While we are not unsympathetic to this view, we presently only want to suggest that folk *attributions* of causation are easily and uncontroversially influenced by biasing factors, and that the law must be alert to this fact.

Given the important legal dimension of our findings, it should be examined whether professional judges are as susceptible to bias in the determination of proximate cause as our lay samples (in particular in civil law jurisdictions, where experts decide the matter).
Finally, scholars working in experimental jurisprudence should investigate whether ordinary judgments of the *actus reus* (also with respect to a number of other problematic effects) could be debiased, and suggest concrete and practicable strategies that common law courts could implement.

8. Conclusion

The Responsibility View and the Bias View come apart in their treatment of factors peripheral to moral responsibility: The former, unlike the latter, predicts that such factors will not influence ordinary causality judgments. In five experiments, we have shown that peripheral factors such as nonpertinent and silly norm violations *do* have a pronounced impact on perceived causation, and that these effects cannot be explained by recourse to potentially legitimate responsibility-enhancing factors such as desire, foreknowledge, or foreseeability. Our results provide considerable evidence in favour of the Bias View, are largely consistent with the Pragmatic View, and they put pressure on the Responsibility View and the Counterfactual View of causation.

The status of the (pertinent) Norm Effect, as it is standardly explored in the literature, requires further examination. According to proponents of the Responsibility View, the effect of pertinent norms can be regarded "peripheral" if it is not mediated by a nonperipheral factors such as negligence or foreseeability. But given that norm-adherence is quite tightly connected to moral responsibility, this stringent criterion could be dropped without the Responsibility View loosing much of its plausibility. (It will still have problems with nonpertinent and silly norms). Naturally, the Bias View, as well as other recent accounts of the Norm Effect, also have plausible explanations of the effect on offer. It thus seems that the debate concerning the status of the Norm Effect might, by now, depend at least as strongly on the plausibility of the theoretical assumptions invoked, as on potential further experimental inquiry.

Acknowledgments

We would like to thank [REDACTED] and three anonymous reviewers for their helpful feedback on the paper. Work on this project was supported by [REDACTED].

References

Alicke, M. D. (1992). Culpable Causation. *Journal of Personality and Social Psychology*, *63*(3), 368–378.

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

Alicke, M. D. (2008). Blaming Badly. *Journal of Cognition and Culture*, *8*(1–2), 179–186.

Alicke, M. D., & Rose, D. (2012). Culpable Control and Deviant Causal Chains. *Personality and Social Psychology Compass*, *6*(10), 723–735.

Alicke, M. D., Rose, D., & Bloom, D. (2011). Causation, Norm Violation, and Culpable Control. *The Journal of Philosophy*, *108*(12), 670–696.

Baron, J. (2008). *Thinking and Deciding* (4th ed.). Cambridge University Press.

Bear, A., & Knobe, J. (2017). Normality: Part Descriptive, Part Prescriptive. *Cognition*, *167*, 25–37.

Bebb, J., & Beebee, H. (2024). Causal Selection and Egalitarianism. In J. Knobe & S. Nichols (Eds.), *Oxford Studies in Experimental Philosophy: Volume 5* (pp. 401–433). Oxford University Press.

Blanchard, T., & Schaffer, J. (2017). Cause Without Default. In H. Beebee, C. Hitchcock, & H. Price (Eds.), *Making a Difference: Essays on the Philosophy of Causation* (pp. 175–214). Oxford University Press.

Bourgeois-Gironde, S. & Kneer, M. (2018). Intention, cause, et responsabilité: Mens Rea et effet Knobe. In Ferey, S. & G'Sell, F. (eds) *Causalité, Responsabilité et Contribution à la Dette*, 117-144.

Brown, T. R. (2023). Minding Accidents. *University of Colorado Law Review*, *94*(1), 89–148.

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., Beebe, J., Berniūnas, R., Boudesseul, J., Colombo, M., Cushman, F., Diaz, R., N'Djaye Nikolai van Dongen, N., Dranseika, V., Earp, B. D., Torres, A. G., Hannikainen, I., Hernández-Conde, J. V., Hu, W., … Zhou, X. (2021). Estimating the Reproducibility of Experimental Philosophy. *Review of Philosophy and Psychology*, *12*(1), 9–44.

Dressler, J. (2015). *Understanding Criminal Law* (7th ed.). LexisNexis.

Engelmann, N. (2022). The role of causal representations in moral judgement [Dissertation]. University of Göttingen. Retrieved from https://ediss.uni-goettingen.de/handle/11858/14231?locale-attribute=de.

Engelmann, N., & Kirfel, L. (2024). Who Caused It? Different Effects of Statistical and Prescriptive Abnormality on Causal Selection in Chains. In K. P. Tobia (Ed.), *The Cambridge Handbook of Experimental Jurisprudence* (forthcoming). Cambridge University Press.

Engelmann, N., & Waldmann, M. R. (2021). A Causal Proximity Effect in Moral Judgment. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*. Retrieved from https://escholarship.org/uc/item/9hp8q72s.

Engelmann, N., & Waldmann, M. R. (2022). How causal structure, causal strength, and foreseeability affect moral judgments. *Cognition*, *226*, 105167.

Frisch, L. K., Kneer, M., Krueger, J. I., & Ullrich, J. (2021). The effect of outcome severity on moral judgement and interpersonal goals of perpetrators, victims, and bystanders. *European Journal of Social Psychology*, *51*(7), 1158-1171.

Gerstenberg, T. (2024). Counterfactual simulation in causal cognition. *Trends in Cognitive Sciences*, *28*(10), 924–936.

Gerstenberg, T., & Icard, T. (2020). Expectations affect physical causation judgments. *Journal of Experimental Psychology: General*, *149*(3), 599–607.

Gill, M., Kominsky, J. F., Icard, T. F., & Knobe, J. (2022). An interaction effect of norm violations on causal judgment. *Cognition*, *228*, 105183.

Giroux, M. E., Coburn, P. I., Harley, E. M., Connolly, D. A., & Bernstein, D. M. (2016). Hindsight bias and law. *Zeitschrift für Psychologie*.

Goldberg, J. C. P., & Zipursky, B. C. (2010). *Torts*. Oxford University Press.

Goulette, V., & Verkampt, F. (2023). Blame-validation: Beyond rationality? Effect of causal link on the relationship between evaluation and causal judgment. *Philosophical Psychology*, online first, 1–20.

Green, L. (1961). Foreseeability in Negligence Law. *Columbia Law Review*, *61*(8), 1401–1424.

Güver, L., & Kneer, M. (2023a). Causation and the Silly Norm Effect. In S. Magen & K. Prochownik (Eds.), *Advances in Experimental Philosophy of Law* (pp. 133–168). Bloomsbury Publishing.

Güver, L., & Kneer, M. (2023b). Causation, Foreseeability, and Norms. In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Meeting of the Cognitive Science Society* (pp. 888–895). Cognitive Science Society.

Hannikainen, I. R., Tobia, K. P., De Almeida, G. D. F., Donelson, R., Dranseika, V., Kneer, M., ... & Struchiner, N. (2021). Are there cross-cultural legal principles? Modal reasoning uncovers procedural constraints on law. *Cognitive Science*, *45*(8), e13024.

Hannikainen, I. R., Tobia, K. P., de Almeida, G. D. F., Struchiner, N., Kneer, M., Bystranowski, P., ... & Żuradzki, T. (2022). Coordination and expertise foster legal textualism. *Proceedings of the National Academy of Sciences*, *119*(44), e2206531119.

Henne, P. (2023). Experimental Metaphysics: Causation. In A. M. Bauer & S. Kornmesser (Eds.), *The Compact Compendium of Experimental Philosophy*. De Gruyter.

Henne, P., & O'Neill, K. (2022). Double Prevention, Causal Judgments, and Counterfactuals. *Cognitive Science*, *46*(5), e13127.

Henne, P., Kulesza, A., Perez, K., & Houcek, A. (2021). Counterfactual thinking and recency effects in causal judgment. *Cognition*, *212*, 104708.

Henne, P., O'Neill, K., Bello, P., Khemlani, S., & De Brigard, F. (2021). Norms Affect Prospective Causal Judgments. *Cognitive Science*, *45*(1), e12931.

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

Henrich, J. (2020). *The WEIRDest people in the world: How the West became psychologically peculiar and particularly prosperous*. Penguin UK.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29-29.

Hitchcock, C., & Knobe, J. (2009). Cause and Norm. *The Journal of Philosophy*, *106*(11), 587–612.

Hsee, C. K. (1996). The Evaluability Hypothesis: An Explanation for Preference Reversals between Joint and Separate Evaluations of Alternatives. *Organizational Behavior and Human Decision Processes*, *67*(3), 247–257.

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

Jaeger, C. B. (2023). Reasonableness from an Experimental Jurisprudence Perspective. *Cambridge Handbook of Experimental Jurisprudence (Kevin Tobia, ed.), Forthcoming*.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, *39*(4), 341–350.

Kamin, K. A., & Rachlinski, J. J. (1995). *Ex post ≠ ex ante*: Determining liability in hindsight. *Law and Human Behavior*, *19*(1), 89–104.

Kirfel, L., & Lagnado, D. (2018). Statistical norm effects in causal cognition. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, *40*, 617–622.

Kirfel, L., & Lagnado, D. (2021a). Causal judgments about atypical actions are influenced by agents' epistemic states. *Cognition*, *212*, 104721.

Kirfel, L., & Lagnado, D. (2021b). Causation by Ignorance. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*, 966–972.

Kirfel, L., & Lagnado, D. (2021c). *Changing Minds — Epistemic Interventions in Causal Reasoning*. PsyArXiv. Retrieved from https://doi.org/10.31234/osf.io/db6ms

Kirfel, L., & Phillips, J. (2021). The Impact of Ignorance Beyond Causation: An Experimental Meta-Analysis. *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*, *43*, 1595–1601.

Kirfel, L., & Phillips, J. (2023). The pervasive impact of ignorance. *Cognition*, *231*, 105316.

Kirfel, L., Bunk, X., Ro'i, Z., & Gerstenberg, T. (2023). Father, don't forgive them, for they could have known what they're doing. *Proceedings of the 45th Annual Conference of the Cognitive Science Society*, *45*, 980–987.

Kneer, M. (2022). Reasonableness on the Clapham Omnibus: Exploring the outcome-sensitive folk concept of *reasonable*. In *Judicial Decision-Making: Integrating Empirical and Theoretical Perspectives* (pp. 25–48). Cham: Springer International Publishing.

Kneer, M., & Bourgeois-Gironde, S. (2017). Mens Rea Ascription, Expertise and Outcome Effects: Professional judges Surveyed. *Cognition*, *169*, 139–146.

Kneer, M., Hannikainen, I., Zehnder, M.-A., Almeida, G., F., A., Bystranowski, P., Dranseika, V., […] and Struchiner, N. (2025). The Severity Effect on Intention and Knowledge: A cross-cultural study with laypeople and legal experts. In preparation.

Kneer, M., & Machery, E. (2019). No luck for moral luck. *Cognition*, *182*, 331–348.

Kneer, M., & Skoczeń, I. (2023). Outcome effects, moral luck and the hindsight bias. *Cognition*, *232*, 105258.

Knobe, J. (2023). Difference and robustness in the patterns of philosophical intuition across demographic groups. *Review of Philosophy and Psychology*, 1-21.

Knobe, J., & Fraser, B. (2008). Causal Judgment and Moral Judgment: Two Experiments. In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Vol. 2* (pp. 441–447). MIT Press.

Knobe, J., & Shapiro, S. (2021). Proximate cause explained. *The University of Chicago Law Review*, *88*(1), 165-236.

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

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2015). Causal superseding. *Cognition*, *137*, 196–209.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(1), 1–26.

Lagnado, D. A. (2021). *Explaining the Evidence: How the Mind Investigates the World*. Cambridge University Press.

Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, *108*(3), 754–770.

Lagnado, D. A., & Gerstenberg, T. (2017). Causation in Legal and Moral Reasoning. In M. R. Waldmann (Ed.), *The Oxford Handbook of Causal Reasoning* (pp. 565–601). Oxford University Press.

Lidén, M., Gräns, M., & Juslin, P. (2019). 'Guilty, no doubt': detention provoking confirmation bias in judges' guilt assessments and debiasing techniques. *Psychology, Crime & Law*, *25*(3), 219-247.

Liu, J. Z., & Li, X. (2019). Legal techniques for rationalizing biased judicial decisions: Evidence from experiments with real judges. *Journal of Empirical Legal Studies*, *16*(3), 630-670.

Livengood, J., & Rose, D. (2016). Experimental Philosophy and Causal Attribution. In J. Sytsma & W. Buckwalter (Eds.), *A Companion to Experimental Philosophy* (pp. 434–449). Blackwell.

Livengood, J., Sytsma, J., & Rose, D. (2017). Following the FAD: Folk Attributions and Theories of Actual Causation. *Review of Philosophy and Psychology*, *8*(2), 273–294.

Macleod, J. (2019). Ordinary Causation: A Study in Experimental Statutory Interpretation. *Indiana Law Journal*, *93*(3), 957–1030.

Margoni, F., & Brown, T. R. (2023). Jurors use mental state information to assess breach in negligence cases. *Cognition*, *236*, 105442.

Margoni, F., & Surian, L. (2022). Judging accidental harm: Due care and foreseeability of side effects. *Current Psychology*, *41*(12), 8774–8783.

Morris, A., Phillips, J., Gerstenberg, T., & Cushman, F. (2019). Quantitative causal selection patterns in token causation. *PLOS ONE*, *14*(8), e0219704.

Murray, S., Krasich, K., Irving, Z., Nadelhoffer, T., & De Brigard, F. (2023). Mental control and attributions of blame for negligent wrongdoing. *Journal of Experimental Psychology: General*, *152*(1), 120.

Nobes, G., & Martin, J. W. (2022). They should have known better: The roles of negligence and outcome in moral judgements of accidental actions. *British Journal of Psychology*, *113*(2), 370–395.

Olier, J. G., Willemsen, P. & Kneer, M. (2025). Ordinary causal attributions, norms, and gradability. In preparation.

Owen, D. (2009). Figuring Foreseeability. *Wake Forest Law Review*, *44*(5), 1277–1308.

Phillips, J., & Knobe, J. (2018). The psychological representation of modality. *Mind & Language*, *33*(1), 65–94.

Phillips, J., Luguri, J. B., & Knobe, J. (2015). Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition*, *145*, 30–42.

Pinillos, N. Á., Smith, N., Nair, G. S., Marchetto, P., & Mun, C. (2011). Philosophy's New Challenge: Experiments and Intentional Action. *Mind & Language*, *26*(1), 115–139.

Prochownik, K. (2022). Causation in the law, and experimental philosophy. In P. Willemsen & A. Wiegmann (Eds.), *Advances in Experimental Philosophy of Causation* (pp. 165–188). Bloomsbury Publishing.

Rachlinski, J. J. (1998). A Positive Psychological Theory of Judging in Hindsight. *The University of Chicago Law Review*, *65*(2), 571–625.

Rachlinski, J. J. (2000). Heuristics and Biases in the Courts: Ignorance or Adaptation? *Oregon Law Review*, *79*(1), 61–102.

Roese, N. J., & Vohs, K. D. (2012). Hindsight Bias. *Perspectives on Psychological Science*, *7*(5), 411–426.

Rogers, R., Alicke, M. D., Taylor, S. G., Rose, D., Davis, T. L., & Bloom, D. (2019). Causal deviance and the ascription of intent and blame. *Philosophical Psychology*, *32*(3), 404–427.

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

Rose, D., & Danks, D. (2012). Causation: Empirical Trends and Future Directions. *Philosophy Compass*, *7*(9), 643–653.

Samland, J., & Waldmann, M. R. (2014). Do social norms influence causal inferences. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual meeting of the cognitive science society* (pp. 1359–1364). Cognitive Science Society.

Samland, J., & Waldmann, M. R. (2015). Highlighting the causal meaning of causal test questions in contexts of norm violations. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th annual meeting of the cognitive science society* (pp. 2092–2097). Cognitive Science Society.

Samland, J., & Waldmann, M. R. (2016). How Prescriptive Norms Influence Causal Inferences. *Cognition*, *156*, 164–176.

Samland, J., Josephs, M., Waldmann, M. R., & Rakoczy, H. (2016). The role of prescriptive norms and knowledge in children's and adults' causal selection. *Journal of Experimental Psychology: General*, *145*(2), 125–130.

Sarin, A., & Cushman, F. (2023). Punishment in negligence is multifactorial: influenced by outcome, lack of due care, and the mere failure of thought. In preparation.

Sarin, A., & Cushman, F. (2024). One thought too few: An adaptive rationale for punishing negligence. *Psychological Review*, *131*(3), 812–824.

Solan, L. M., & Darley, J. M. (2001). Causation, Contribution, and Legal Liability: An Empirical Study. *Law and Contemporary Problems*, *64*(4), 265–298.

Spamann, H., & Klöhn, L. (2016). Justice is less blind, and less legalistic, than we thought: Evidence from an experiment with real judges. *The Journal of Legal Studies*, *45*(2), 255-280.

Stich, S. P., & Machery, E. (2023). Demographic differences in philosophical intuition: A reply to Joshua Knobe. *Review of Philosophy and Psychology*, *14*(2), 401-434.

Stratenwerth, G., & Kuhlen, L. (2011). *Die Straftat* (6th ed.). Vahlen.

Strohmaier, N., Pluut, H., Van den Bos, K., Adriaanse, J., & Vriesendorp, R. (2021). Hindsight bias and outcome bias in judging directors' liability and the role of free will beliefs. *Journal of Applied Social Psychology*, *51*(3), 141-158.

Summers, A. (2018). Common-Sense Causation in the Law. *Oxford Journal of Legal Studies*, *38*(4), 793–821.

Sytsma, J. (2019a). Structure and norms: Investigating the pattern of effects for causal attributions [Preprint]. Retrieved from http://philsci-archive.pitt.edu/16626/.

Sytsma, J. (2019b). The Character of Causation: Investigating the Impact of Character, Knowledge, and Desire on Causal Attributions [Preprint]. Retrieved from http://philsci-archive.pitt.edu/16739/.

Sytsma, J. (2021). Causation, Responsibility, and Typicality. *Review of Philosophy and Psychology*, *12*(4), 699–719.

Sytsma, J. (2022). The Responsibility Account. In P. Willemsen & A. Wiegmann (Eds.), *Advances in Experimental Philosophy of Causation* (pp. 145–164). Bloomsbury Publishing.

Sytsma, J., Livengood, J., & Rose, D. (2012). Two Types of Typicality: Rethinking the Role of Statistical Typicality in Ordinary Causal Attributions. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(4), 814–820.

Sytsma, J., Willemsen, P., & Reuter, K. (2023). Mutual entailment between causation and responsibility. *Philosophical Studies*, *180*(12), 3593–3614.

Tobia, K. P. (2018). How people judge what is reasonable. *Ala. L. Rev.*, *70*, 293.

Tobia, K. (2021). Law and the Cognitive Science of Ordinary Concepts. In B. Brozek, J. Hage, & N. Vincent, *Law and Mind: A Survey of Law and the Cognitive Sciences* (pp. 86–96). Cambridge University Press.

Tobia, K., Hannikainen, I.R., Kamper, D., Almeida, G., Strohmaier, N., Dranseika, V.,

Kneer, M. […], Struchiner, N. (2025). What is reaonsable? A multi-country study. *Stanford Journal of International Law* (forthcoming).

VerSteeg, R. (2011). Perspectives on Forseeability in the Law of Contracts and Torts: The Relationship between Intervening Causes and Impossibility. *Michigan State Law Review*, *2011*, 1497–1519.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

Willemsen, P., & Kirfel, L. (2019). Recent empirical work on the relationship between causal judgements and norms. *Philosophy Compass*, *14*(1), e12562.